UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1995

SUMMER FACULTY RESEARCH PROGRAM FINAL REPORTS

VOLUME 3B

PHILLIPS LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major David Hart

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Program Administrator
Johnetta Thompson

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1995

19981218 058

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-98-

0823

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE December, 1995	3. REPOR. Final

4. TITLE AND SUBTITLE
USAF Summer Research Program - 1995 Summer Faculty Research Program Final Reports, Volume 3B, Phillips Laboratory

5. FUNDING NUMBERS

6. AUTHORS
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research and Development Labs, Culver City, CA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
4040 Fairfax Dr, Suite 500
Arlington, VA 22203-1613

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Contract Number: F49620-93-C-0063

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Faculty Research Program (USAF- SFRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory Technical Directorates and Air Force Air Logistics Centers. Each participant provided a report of their research, and these reports are consolidated into this annual report.

14. SUBJECT TERMS
AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, SUMMER, UNIVERSITIES

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

PREFACE

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

Due to its length, Volume 3 is bound in two parts, 3A and 3B. Volume 3A contains #1-21, and Volume 3B contains reports #22-39. The Table of Contents for Volume 3 is included in both parts.

This document is one of a set of 16 volumes describing the 1995 AFOSR Summer Research Program. The following volumes comprise the set:

VOLUME	TITLE
1	Program Management Report
	Summer Faculty Research Program (SFRP) Reports
2A & 2B	Armstrong Laboratory
3A & 3B	Phillips Laboratory
4	Rome Laboratory
5A, 5B, & 5C	Wright Laboratory
6A & 6B	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	Graduate Student Research Program (GSRP) Reports
7A & 7B	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10A & 10B	Wright Laboratory
11	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	High School Apprenticeship Program (HSAP) Reports
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

SFRP FINAL REPORT TABLE OF CONTENTS   i-xiv

SFRP FINAL REPORTS

On the Scintillation of Transionospheric Signals

S. P. Kuo

Professor

Department of Electrical Engineering

Polytechnic University

Route 110

Farmingdale, NY 11735

Final Report for:

Summer Research Projects

Geophysics Directorate of the Phillips Laboratory

September, 1995

On the Scintillation of Transionospheric Signals

S. P. Kuo

Professor

Department of Electrical Engineering

Polytechnic University

Abstract

The spectral effect of equatorial ionospheric irregularities on the scintillation of transionospheric signals is considered. A two dimensional periodic density modulation is used to model the bottomside sinusoidal (BSS) irregularity structure found in the equatorial F region. A quasi-particle method is introduced to examine the effect of the spectral width Δk of the density modulation on wave scattering. It is exemplified by calculating the dependence of the scintillation index S_4 of the transionospheric electromagnetic signal on Δk. The parametric dependences of S_4 on the longitudinal and transverse scale lengths of the density irregularities are also presented. The results show that the optimum S_4 depends strongly on $\Delta k / k$. S_4 decreases with $\Delta k / k$ to a minimum, therewith exhibiting a damped oscillation. At the minimum, S_4 reduces to a negligibly small value almost independent of the average scale length of the modulation when the longitudinal scale length of the irregularity is allowed to vary and the transverse irregularity is fixed.

On the Scintillation of Transionospheric Signals

S. P. Kuo

INTRODUCTION

The scintillation of the transionospheric signal is attributed to reflection and scattering processes experienced by the wave propagating in an inhomogeneous ionosphere. When the inhomogeneity has a periodic pattern, scattering becomes the most effective disturbance to wave propagation, even if the Bragg Scattering condition is not satisfied. This is because the accumulating result of multiple scattering by the repeating density fluctuations in the medium can significantly perturb the amplitude and phase of the propagating wave. Apparently, the significance of multiple scattering depends on the magnitude of the perturbation in the ionosphere. In general, the level of density fluctuation can be measured in terms of a Scintillation Index, S_4. Normally, S_4 is small when the plasma is only perturbed by random fluctuations. On the other hand, it has been shown that the scintillation of a 250 MHz Beacon Satellite signal when passing through the equatorial ionosphere could vary between a moderate level $S_4 = 0.1$ and a fairly strong level $S_4 = 0.8$. During the measurements, it was also observed by the same Atmosphere Explorer Satellite that the ionosphere was perturbed by large amplitude plasma irregularities, characterized by nearly sinusoidal waveforms [Basu et al., 1986]. The waveforms extend in long-lived patches, for a few orbital periods (~ 100 minutes), more than 7000 km in the E-W directions in a narrow belt extending approximately $\pm 12°$ from the dip equator on the bottomside of the F region [Valladres et al., 1983].

When the multiple scattering process becomes important, the conventional approximations used in wave propagation study, such as WKB and forward scattering, are no longer valid. More sophisticated approaches such as the Phase Screen [Wagen and Yeh, 1989], and Rytov Approximations [Keller, 1969], are often restricted to the weak scintillation cases. Therefore, a quasi-particle method to deal with multiple scattering in wave propagation was developed recently [Wu and Marcuvitz, 1983: Ho et al., 1994]. The wave is considered as a group of quasi-particles following a Wigner distribution function [Bremmer, 1973] in the $(\vec{k}, \vec{r})$ space, which is governed by a kinetic equation with an effective collision term corresponding to the attenuation and scattering of the wave. This kinetic equation leads to a hierarchy of moment equations in the $\vec{r}$ space. It has been shown [Ho et al., 1994] that this system of equations can be truncated to the second moment for a cold quasi-particle distribution, as the case of a monochromatic satellite signal. The agreement between the results of the previous analysis and the existing experimental data on the scintillation of Atmosphere Explorer Satellite signals [Basu et al., 1986] provides evidence of the advantage of the quasi-particle method over other methods in dealing with wave propagation in a strong multiple scattering environment.

In the present work, the previously developed quasi-particle method is employed to study how the results of multiple scattering are affected by spectral width $\Delta k / k$ of the irregularities. The S_4 is optimized by matching the longitudinal and transverse scale lengths of the irregularity to the corresponding Fresnel size. In the following, a review of quasi-particle formulation is presented. The effective collision term is then extended for irregularities having a finite spectral width. The numerical results including the dependences of S_4 on the spectral width and scale length of the

irregularity will then be presented and discussed. The conclusion of the work will finally be drawn.

FORMULATION

The wave characteristics including its amplitude and phase can be described by a joint distribution function of the position vector $\vec{r}$ and the wavevector $\vec{k}$, known as the **Wigner Distribution Function** which is defined to be

$$F(\vec{r},\vec{k},t) = \pi^{-2} \int d\vec{r}' \exp(2i\vec{k} \cdot \vec{r}') \, \psi^{*}(\vec{r}+\vec{r}',t) \, \psi(\vec{r}-\vec{r}',t) \tag{1}$$

where ψ is a wave field component and the integration is over all space where the field exists. The quasi-particle spatial density and spectral density can be obtained from the Wigner distribution function by the two relations $|\psi(\vec{r},t)|^2 = \int F(\vec{r},\vec{k},t) d\vec{k}$ and $|G(\vec{k},t)|^2 = \int F(\vec{r},\vec{k},t) d\vec{r}$, respectively.

Consider a modeled single-mode wave equation expressed in the form as

$$i\frac{\partial \psi}{\partial t} = -\left(\nabla^2 + k_0^2 \varepsilon_1(\vec{r})\right)\psi \tag{2}$$

The corresponding transport equation of the Wigner distribution function is derived to be [Ho et al., 1994]:

$$\vec{k} \cdot \nabla F = -(k_0^2 / \pi^2) \int d\vec{k}' F(\vec{r},\vec{k}') \int d\vec{s} \{ \varepsilon_{1r}(\vec{r}-\vec{s}) \sin\left[2(\vec{k}-\vec{k}') \cdot \vec{s}\right]$$
$$+ \varepsilon_{1i}(\vec{r}-\vec{s}) \cos\left[2(\vec{k}-\vec{k}') \cdot \vec{s}\right] \} \tag{3}$$

where $\varepsilon_1(\bar{r}) = \varepsilon(\bar{r}) - \varepsilon_0 = \varepsilon_{1r}(\bar{r}) + i\varepsilon_{1i}(\bar{r})$ and ε_0 is the permittivity of the unperturbed background.

We now consider the wave scattering within a layer of plasma density irregularities represented by a two dimensional variation of plasma density which has the form

$$n(x,y) = P_\ell(x)\{n_0 + \Delta n[1 - (\sin k_{11}x - \sin k_{10}x)/\Delta k_1 x]$$
$$[1 - (\sin k_{21}\eta - \sin k_{20}\eta)/\Delta k_2 \eta]\}$$

where $\eta = \cos\theta(y + x\tan\theta)$, $P_\ell(x) = 1$ for $|x| \le \ell$ and 0 for $|x| \ge \ell$ representing a gate function. This density variation can be decomposed into two separate components. The first is characterized by $\langle k_1 \rangle = (k_{11} + k_{10})/2 = 2\pi/d_1$ with the average periodicity d_1 in the longitudinal direction. The second component has an average periodicity d_2, characterized by $\langle k_2 \rangle = (k_{21} + k_{20})/2 = 2\pi/d_2$, oblique to the first with an angle $\phi = \pi/2 - \theta$; n_0 is the background plasma density, and Δn is the amplitude of the density irregularity.

This variation consists of two gratings offset from each other by an angle ϕ as illustrated in Fig. 1. As shown these two gratings extend to infinity in the transverse direction and have a finite extent 2ℓ along the downward propagation direction.

It is expected that a model provide a description of irregularities and not resulting signal scintillation exclusively [Fremouw et al., 1984]. This model provides a useful description of BSS irregularities containing information on irregularity shape as well as strength. In addition parameterization allows adjustment due to geophysical changes.

The complex dielectric function is derived from the plasma density function as

$$\varepsilon(x,y) = 1 - [\omega_{p0}^2/(\omega^2 + v^2)](1 - iv/\omega)P_\ell(x)\{1 + \beta[1 - \delta_1(x)][1 - \delta_2(\eta)]\}$$

$$= \varepsilon_0 + \varepsilon_1(x,y) \tag{4}$$

where $\delta_i(\xi) = \left(\sin k_{i1}\xi - \sin k_{i0}\xi\right)/\Delta k_i\xi$ $i=1,2$; $\varepsilon_0 = 1 - \alpha P_\ell(x)(1 - iv/\omega)$,

$\varepsilon_1 = -\beta\alpha P_\ell(x)(1 - iv/\omega)[1 - \delta_1(x)][1 - \delta_2(\eta)]$, $\omega_p/2\pi$ is the background electron plasma frequency, ω is the wave angular frequency, v is the electron-ion collision frequency; $\beta = \Delta n/n_0$ is the percentage of density fluctuation with respect to the background plasma density; and $\alpha = \omega_p^2/(\omega^2 + v^2) \ll 1$ is assumed.

The right hand side (RHS) of the transport equation (3) is integrated once to become

$$\left[(k_0 + k_x)\frac{\partial}{\partial x} + k_y\frac{\partial}{\partial y}\right]F = \beta\alpha(k_0^2/\pi)G_1 \otimes F \tag{5}$$

where

$$G_1 = \delta(2k_y)g_1(k_x)h(k_x x) - (1/\Delta k_2 \cos\theta)g_1(k_x + k_y \tan\theta)$$

$$h(k_x x + k_y y)[U(k_y - \bar{k}_{20}) - U(k_y - \bar{k}_{21})]$$

$$g_1(k) = (2\sin 2k\ell/k) - (2/\Delta k_1)\sum_{n=0}^{\infty}(-1)^n[(2\ell)^{2n+1}/(2n+1)(2n+1)!]$$

$$\{(k - k_{10}/2)^{2n+1} + (k + k_{11}/2)^{2n+1} - (k - k_{11}/2)^{2n+1}$$

$$-(k + k_{10}/2)^{2n+1}\}$$

$h(\xi) = \sin 2\xi - (v/\omega)\cos 2\xi$, $\bar{k}_{ij} = (1/2)k_{ij}\cos\theta$, $\otimes$ stands for convolution; with $k_{i0} \leq k_i \leq k_{i1}$ $i=1,2$; $\Delta k_1 = k_{11} - k_{10}$ and $\Delta k_2 = k_{21} - k_{20}$; δ is the Dirac delta function.

In calculating the scintillation index:

$$S_4 = \left([<\rho^2> - <\rho>^2] / <\rho>^2 \right)^{1/2},$$

where the average is performed over a spatial period of the transverse irregularity, only the intensity distribution $\rho = |\psi(\bar{r},t)|^2$ of the wave is needed. Thus, the moment equations of (5) are employed in the following analysis. The first three moments of the quasi-particle distribution function are: the quasi-particles' number density, current density and energy density, which are defined respectively to be $\rho = \int d\bar{k} F(\bar{r},\bar{k})$, $\bar{J} = \int d\bar{k}\bar{k}F(\bar{r},\bar{k}) = \bar{k}_0\rho + \bar{J}_1$, and

$$\underline{s} = \int d\bar{k}\bar{k}\bar{k}F(\bar{r},\bar{k}) = \bar{k}_0\bar{k}_0\rho + \bar{k}_0\bar{J}_1 + \bar{J}_1\bar{k}_0 + \underline{\underline{s}}_1.$$

The first two moment equations of (5) are found to be

$$k_0\frac{\partial}{\partial x}\rho + \nabla\cdot\bar{J}_1 = \beta\alpha(k_0^2/\pi)G\rho$$

and

(6)

$$k_0\frac{\partial}{\partial x}\bar{J}_1 + \nabla\cdot\underline{\underline{s}}_1 = \beta\alpha(k_0^2/\pi)[\bar{H}\rho + G\bar{J}_1]$$

where

$$G(\bar{r}) = -\pi(v/\omega)P_\ell(x)[1-\delta_1(x)][1-\delta_2(\eta)]$$

and

$$
\begin{aligned}
\bar{H}(\bar{r}) = \hat{x}\, \tfrac{\pi}{2} \langle & \{[\delta(x-\ell)-\delta(x+\ell)][1-\delta_1(\ell)]+(1/\Delta k_1 x)P_\ell(x) \\
& [k_{11}\cos k_{11}x - k_{10}\cos k_{10}x - \Delta k_1 \delta_1(x)]\}[1-\delta_2(\eta)] \\
& -(\tan\theta/\Delta k_2 \eta)P_\ell(x)[1-\delta_1(x)][k_{21}\cos k_{21}\eta \\
& -k_{20}\cos k_{20}\eta - \Delta k_2 \delta_2(\eta)]\rangle + \hat{y}\,\tfrac{\pi}{2}(1/\Delta k_2 \eta)P_\ell(x) \\
& [1-\delta_1(x)][k_{21}\cos k_{21}\eta - k_{20}\cos k_{20}\eta - \Delta k_2 \delta_2(\eta)]
\end{aligned}
$$

To truncate the hierarchy of the moment equations, a cold quasi-particle distribution is assumed. This assumption is based on the fact that the initial quasi-particle distribution has zero temperature of the beacon signal. Thus, the second moment tensor $\underline{\underline{s_1}}$ becomes zero and the two moment equations in (6) form a closed set of equations for further analysis.

NUMERICAL ANALYSIS

In the bottomside of the equatorial F region, irregularities having nearly sinusoidal waveforms in the plasma density were observed and extended up to 7000 km in the E-W directions. The rms irregularity amplitude was about 3.5% of the background electron density, which was about $9.9x10^{10}m^{-3}$. The scintillation index S_4 on 250 MHz satellite signal varied from 0.1 to 0.8. The thickness of the irregularity layer was about 50 km. The satellite was located at an altitude of about 400 km and the slant range of the signal was about 600 km. The background ionosphere drifts at an approximate speed of 100 m/s.

The previous work assumed that the irregularity has zero line width, with $500m \leq d_1 \leq 2000m$, $d_2 = 1 km$, and $0 \leq \phi \leq 90°$ to model a nearly sinusoidal waveform. The calculated values of S_4 from a set of equations similar to (6), agreed well with those measured by Basu et al. [1986]. The scale length around 1 km was chosen to obtain 0.1 Hz scintillation for a 100m/s drift as observed by Basu et al. [1986].

In the present work, the effect of the spectral width $\Delta k / k$ of the density irregularities on the intensity distribution of the transionospheric signal in terms of the scintillation index S_4 at the exit of the layer is first examined for 250MHz beacon signal, where $\Delta k_1/ \langle k_1 \rangle = \Delta k_2/ \langle k_2 \rangle = \Delta k / k$ is assumed. The optimum S_4 is evaluated by choosing d_1 and d_2 corresponding to the Fresnel size.

Considering three different cases, $d_1 = 990$ m, 500 m, and 1500 m, and fixing $d_2 = 990$ m and $\phi=45°$, the dependence of the optimum S_4 on spectral width $\Delta k / k$ is evaluated. Presented in Fig. 2 is the result of this optimum S_4 ($d_1 = 990$ m, $d_2 = 990$ m) together with two other cases evaluated with

different d_1. It is shown that the location and value of the first peak of the S_4 curve varies with d_1. A small spectral width in the irregularity spectrum reduces the optimized S_4 value, but enhances S_4 value in other cases assumed. However, the tendency of the effect of spectral width is to reduce the S_4 value. It shows that in all cases presented the S_4 value drop to a minima near $\Delta k / k \cong 0.1$. These minimum locations are almost independent of d_1 as shown in Fig. 3, which plots the dependence of S_4 on d_1 for $\Delta k / k \cong 0.1$. As $\Delta k / k$ increases further above 0.1, S_4 curves as shown in Fig. 2 are featured with damped oscillation. The peaks and minimum appear at or near the same locations, independent of d_1. Since the values of S_4 in the region of $\Delta k / k > 0.1$ are quite small, $\Delta k / k < 0.1$ will be the region of interest. The diminishing effect of the spectral width on the S_4 value suggests that the spectral distribution of the irregularities have smeared out the result of the multiple scattering and hence reduced the net scintillation effect on the transverse signals.

We now fix $d_1 = 990$ m, and $\phi = 45°$, the dependence of S_4 on spectral width $\Delta k / k$ for $d_2 = 990$ m, 1500 m, and 2000 m are evaluated and presented in Fig. 4. The value of the peak of S_4 does not vary with d_2 and the peaks are located at zero spectral width in all cases. The spectral width reduces the S_4 value very quickly for values of d_2 greater than 1000 m. The damped oscillation exhibited previously when $d_2 = 990$ m is not existent in higher values of d_2. In other words, $\Delta k / k \cong 0.1$ is not a minima point when $d_2 > 990$ m.

CONCLUSION

The effect of ionospheric irregularities on the propagation of beacon satellite signals through the ionosphere is studied. A quasi-particle approach is adopted for the analysis. The present work extends the previous one by considering irregularities having a finite spectral bandwidth Δk.

The results show that the Scintillation Index S_4 depends strongly on $\Delta k / k$ and signal frequency. With about 10% spectral width, the scintillation index drops to a minimum having a very small value and almost independent of the longitudinal scale lengths d_1 of the irregularities. However, in the region of small $\Delta k / k$ (<10%), the spectral width of the irregularities may reduce or enhance the scintillation level, depending on the longitudinal scale length d_1 of the irregularities. The overall effect of the spectral distribution is to make the dependence of the scintillation level less sensitive to the scale length of the irregularity.

The results of the present study suggest that the information on scintillation level alone does not provide the actual condition of the ionosphere . A large S_4 value certainly indicates the presence of a high level of irregularities in the ionosphere, on the other hand, a small S_4 value could be the result of a low level of irregularities, or irregularities with a large spectral width.

This uncertainty may be resolved by measuring S_4 simultaneously using two different beacon signals with a large frequency difference. The other approach which elevates the need for the use of two different signals is adding a measurement of phase scintillation. Applying the present theory for the study of the phase scintillation as well as the dependence of S_4 on signal frequency will be presented in our future work.

REFERENCES

Basu, S., Eilleen Martin MacKenzine, Su. Basu, E. Costa, P. Fougere, H. C. Carlson, Jr., H. E. Whitney, 250MHz/GHz Scintillation Parameters in the Equatorial, Polar, and Auroral Environments, IEEE Journal on selected areas in Communications, SAC-5(2), 102-115, 1987

Basu, S., Su. Basu, C. E. Valladares, A. Dasgupta, and H. E. Whitney, Scintillations Associated With Bottomside Sinusoidal Irregularities in the Equatorial F region, J. Geophys. Res., 91(A1), 270-276, 1986.

Bremmer, H., General remarks concerning theories dealing with scattering and diffraction in random media, Radio Sci., 8(6), 511-534, 1983.

Fremouw, E. J., and J. A. Secan, Modeling and scientific application of scintillation results, Radio Sci., 19(3), 687-694, 1984

Ho, Antony Y., S. P. Kuo, and M. C. Lee, Analysis of Electromagnetic Wave Scattering by Ionospheric Irregularities, Radio Sci., 29(5), 1179-1186, 1994.

Keller, J. B., Accuracy and validity of Born and Rytov approximation, J. Opt. Soc. Amer., 59,1003-1004, 1969

Valladares, C. E., W. B. Hanson, J. P. McClure, and B. L. Cragin, Bottomside Sinusoidal Irregularities in the Equatorial F Region, J. Geophys. Res., 88(A10), 8025-8042, 1983

Wagen, J.-F., and K.C. Yeh, Simulation of HF propagation and angle of arrival in a turbulent ionosphere, Radio Sci., 24(2), 196-208, 1989.

Wu, D. M., and N. Marcuvitz, Ionospheric scintillations, Radio Sci., 18(4), 589-607, 1983.
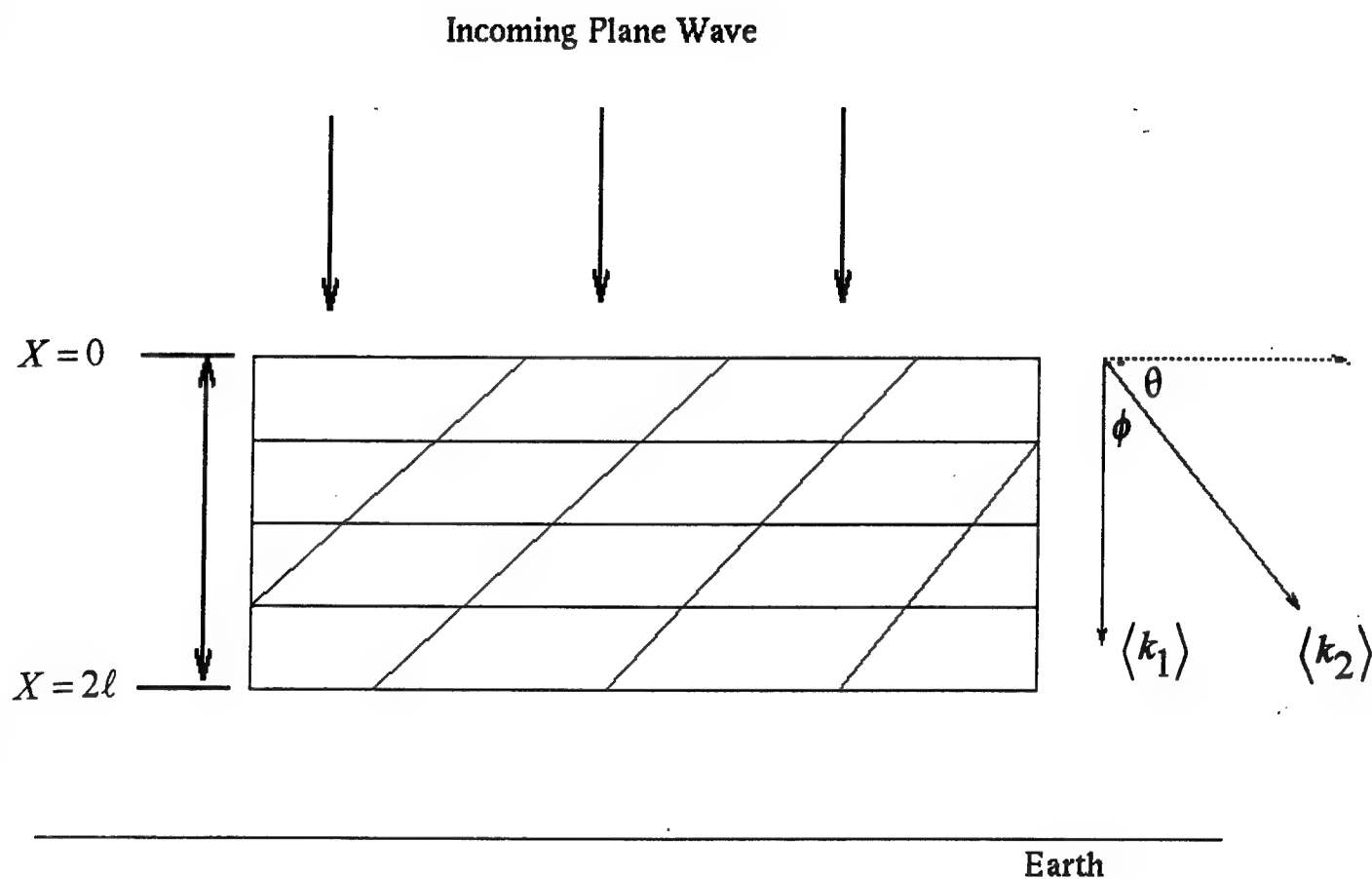
Fig. 1. Incoming plane wave is scattered by a two-dimensional irregularity grating characterized with grating wavevectors $\langle k_1 \rangle$ and $\langle k_2 \rangle$ offset by angle ϕ. The thickness of the irregularity region is 2ℓ.
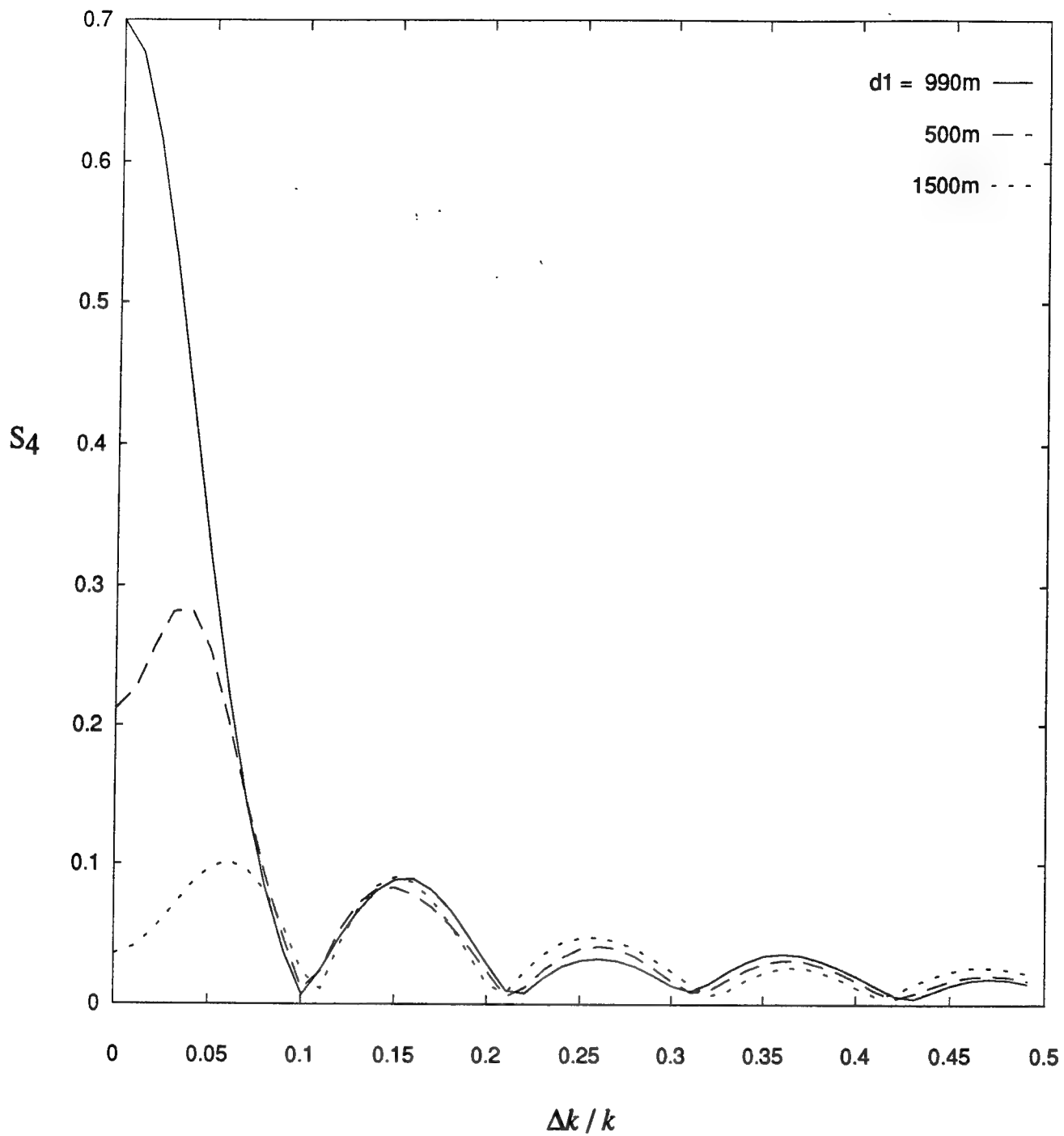
Fig. 2.   Dependence of S_4 on spectral width $\Delta k / k$ for 250MHz beacon signal, with $d_1 = 990$ m, 500 m, and 1500 m, $d_2 = 990$ m, $2\ell = 50$ km, $n_0 = 1 \times 10^{11} m^{-3}$, $\phi = 45°$, and $\beta = 3.5\%$, where scale lengths correspond to optimized S_4.
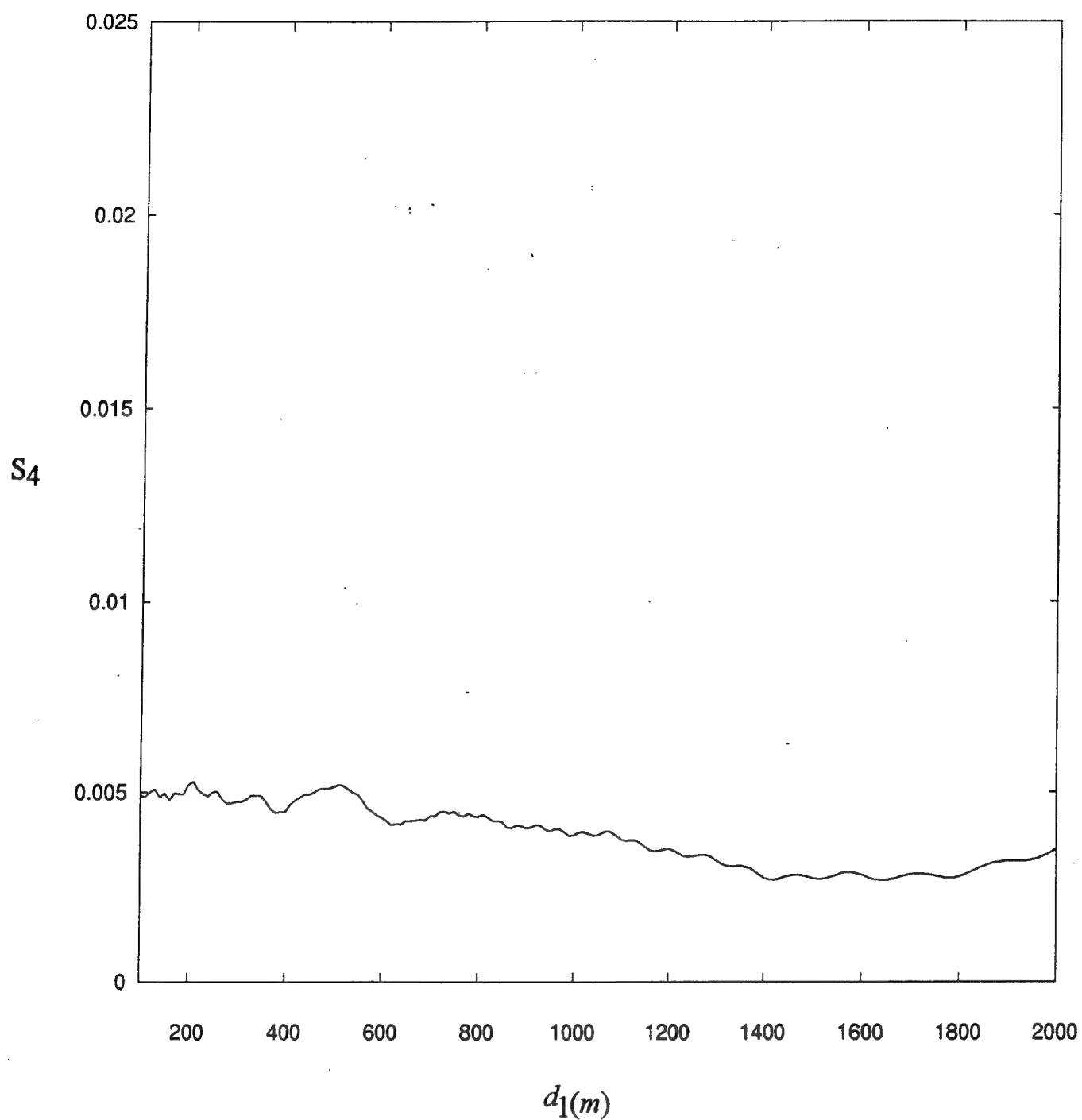
Fig. 3.   Dependence of S_4 on d_1 for 250MHz beacon signal, with
$\Delta k / k \cong 0.1$, $d_2 = 990$ m, $2\ell = 50$ km, $n_o = 1 \times 10^{11} m^{-3}$, $\phi = 45°$
and $\beta = 3.5\%$.
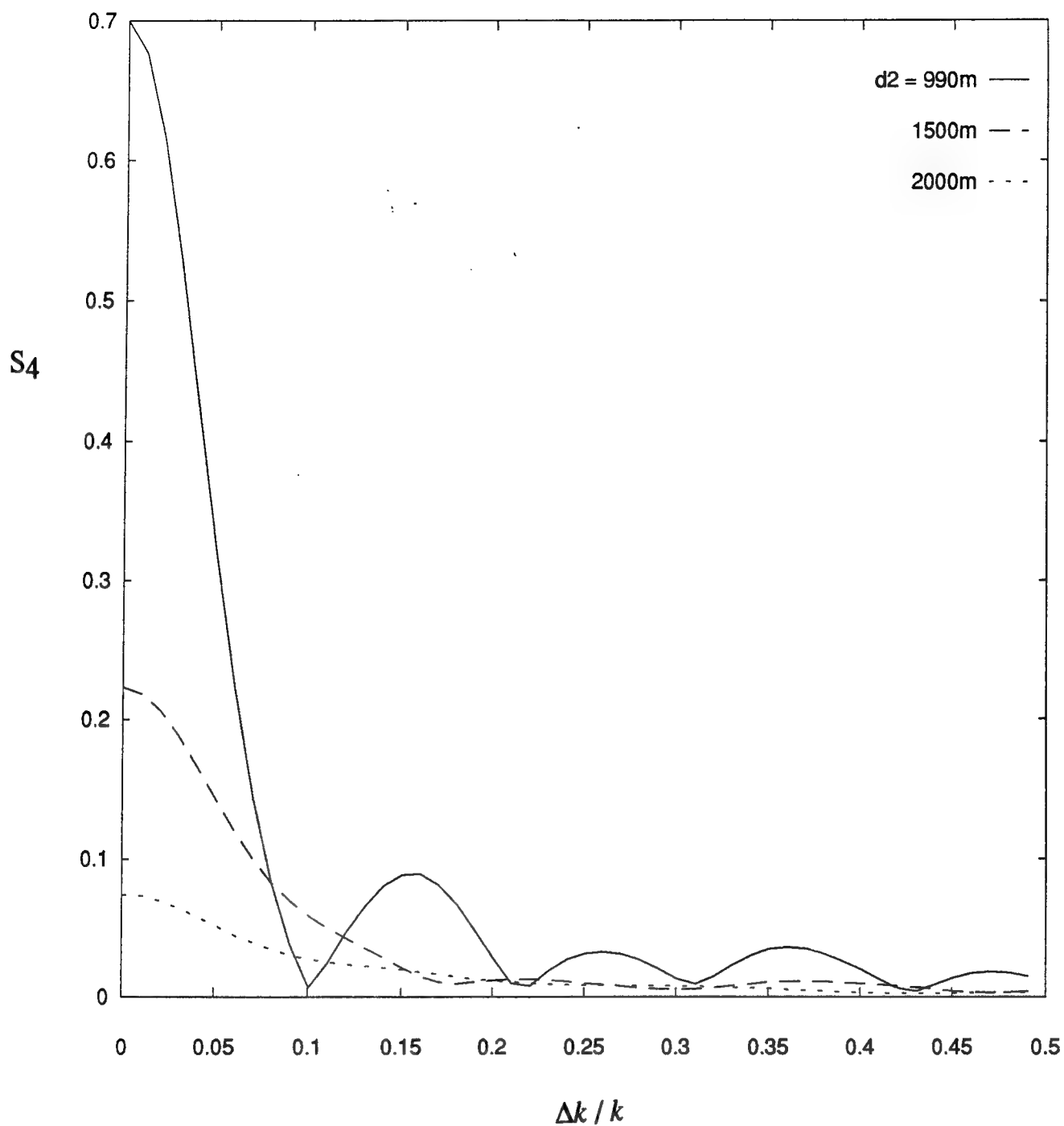
Fig. 4.   Dependence of S_4 on spectral width $\Delta k / k$ for 250MHz beacon signal, with $d_2 = 990$ m, 1500 m, and 2000 m, $d_1 = 990$ m, $2\ell = 50$ km, $n_o = 1 \times 10^{11} m^{-3}$, $\phi = 45°$, and $\beta = 3.5\%$, where scale lengths correspond to optimized S_4.

ACKNOWLEDGMENTS

MATHEMATICAL MODELING FOR THE
THERMIONIC - AMTEC CASCADE

M.A.K. Lodhi
Professor
Department of Physics and Engineering Physics

Texas Tech University
Lubbock, Tx 79409

MATHEMATICAL MODELING FOR THE
THERMIONIC-AMTC CASCADE

M.A.K. Lodhi
Professor
Department of Physics and Engineering Physics
Texas Tech University

Abstract

A Mathematical modeling of a cascade consisting of a TIEC and AMTEC has been performed. The cascade has been broken down in nine nodes. With boundary conditions determined at those nodes, nine simultaneous nodal equations have been constructed and solved by MATHCAD 5.0. View factors for the geometry of the cascade have been calculated and inserted in the appropriate places in the nodal equations. This is done for the first time to our knowledge for this cascade modeling. This effort makes the model relatively more realistic.

MATHEMATICAL MODELING FOR THE
THERMIONIC - AMTEC CASCADE

M.A.K. Lodhi

Introduction

Thermal or mechanical energy-to-electrical energy conversions have been historically the mainstay for power systems. Historically, with the advent of Faraday's Law of electromagnetic induction and steam engines, the dynamic conversion systems evolved with a rapid pace, and were perfected with engineering details mostly for the earth and air base devices. For space power systems, however, as a potential alternative to dynamic conversion systems, static thermal-to-electronic conversion systems have been investigated. With the discovery of Seeback effect the thermoelectromotive force, which occurs in a material or materials under the effect of a temperature gradient, has been the driving force of the thermoelectric generators. The goal for static conversion systems has been to identify static converters with efficiencies that are competitive with dynamic systems. In that a large number of devices have been evolved and studied. Of these the thermionic energy conversion (TIEC), the thermoelectric energy conversion (TOEC), the thermo-photovoltaic converter (TPVC), the alkali metal thermoelectric conversion (AMTEC), the hydrogen thermoelect converter (HYTEC), the thermoacoustic power conversion (TAPC), and the Liquid Metal magnetohydradynamic (LMMHD) are the principal technologies that have been studied. The salient features of these systems are summarized in Table I. The effectiveness of these devices is compared with respect to a number of aspects such as the cost, the load, the space occupancy (area and volume both per energy unit output), temperature range, power range, heat source, lifetime, the present and projected technological status. Comparing these features in Table I, AMTEC seems to be the most mature one of these devices. Several key fundamental issues have been resolved. Engineering issues are presently being addressed. The goal is to identify a heat emitter-converter combination that would enable the system with efficiencies at least 30% and more. In this scenario, for AMTEC to work, it has to be cascaded with a thermal-to-electrical energy converter device. Such a cascade is presently under the investigation at the Air Force Phillips Lab (AFPL or just PL) under the leadership of Michael Schuller. This cascade consists of TIEC and AMTEC. The TIEC is heated by an electron bombardment which converts part of this bombarded energy into electricity by the

TIEC which is supplied to the load and rest of the energy is rejected. AMTEC picks up this rejected heat from TIEC and converts into electrical energy. A schematic diagram of such a cascade is given in Fig 1.

Cascade Modeling

For the purpose of keeping the track of the heat and mass flow the cascade has been broken down into nine different nodes. At each node the boundary conditions are determined for the heat and mass flow through the cascade. This gives rise to nine simultaneous equations with nine temperatures at these nodes to be solved. A nodal circuit diagram given by Fig 2 is an indicative of heat and mass flow at each node. With the help of the nodal circuit the nine equations are developed and solved for those temperatures. In order to solve these lengthy and complicated equations and iterative procedure is adopted with MATHCAD 5.0. For the iterative solution some initial values of the unknowns are guessed and supplied for the calculations. The final answers do not depend on the initial values supplied. The answers do depend on the parameters used for material properties and radiation loss or heat transfer terms. The radiation terms are in turn dependent on view factors of the geometry of the cascade.

View Factors

Before solving the nine nodal equations it is essential to estimate the view factors at each node as close as possible to the geometry of the cascade. This is not a trivial job. Various view factors have been evaluated and substituted in the appropriate places in the nodal equations with corresponding portions of the areas of the various components of the cascade. The details of calculation of the view factors is avoided here to be presented for the sake of the brevity. The development of these expressions are given as the input and output data of the MATCHED compilation at the outset of the nine simultaneous nodal equations and their solutions.

Nodal Equations

In the next ten pages the nodal equations and their solutions are given. The first few pages are devoted in developing the expressions for calculating view factors. Again for the brevity a nomenclature of the symbols list has been avoided. The nomenclatures are mostly self explanatory and the parameters defined in terms of the others are sequential and some may not have any physical meaning.

$r_1 := 1 \cdot cm$ $r_2 := 1.5 \cdot cm$ $r_3 := 2.15 \cdot cm$ $r_4 := 1.9 \cdot cm$ $r_5 := 3.75 \cdot cm$ $i := 1 .. 10$

$j := 1 .. 10$

$k := 1 .. 8$

$r_6 := r_4$ $r_7 := 2.65 \cdot cm$ $r_8 := 1.525 \cdot cm$ $r_9 := 4.925 \cdot cm$ $r_{10} := 2.85 \cdot cm$

$l_1 := .8 \cdot cm$ $l_2 := .4 \cdot cm$ $l_3 := .2 \cdot cm$ $l_4 := .7 \cdot cm$ $l_5 := 1.2 \cdot cm$ $r_j :=$ $r_i :=$ $l_k :=$

$l_6 := 1 \cdot cm$

$R := \dfrac{r_2}{r_1}$ $L := \dfrac{l_1}{r_1}$ $L_{i,k} := \dfrac{l_k}{r_i}$

r_j	r_i	l_k
1	1	.8
1.5	1.5	.4
2.15	2.15	.2
1.9	1.9	.7
3.75	3.75	1.2
1.9	1.9	1
2.65	2.65	.9
1.525	1.525	3.3
4.925	4.925	
2.85	2.85	

$R_{i,j} := \dfrac{r_j}{r_i}$ $r := r \cdot cm$ $l := l \cdot cm$

$R =$

	1	2	3	4	5	6	7	8	9
1	1	1.5	2.15	1.9	3.75	1.9	2.65	1.525	4.925
2	0.667	1	1.433	1.267	2.5	1.267	1.767	1.017	3.283
3	0.465	0.698	1	0.884	1.744	0.884	1.233	0.709	2.291
4	0.526	0.789	1.132	1	1.974	1	1.395	0.803	2.592
5	0.267	0.4	0.573	0.507	1	0.507	0.707	0.407	1.313
6	0.526	0.789	1.132	1	1.974	1	1.395	0.803	2.592
7	0.377	0.566	0.811	0.717	1.415	0.717	1	0.575	1.858
8	0.656	0.984	1.41	1.246	2.459	1.246	1.738	1	3.23
9	0.203	0.305	0.437	0.386	0.761	0.386	0.538	0.31	1
10	0.351	0.526	0.754	0.667	1.316	0.667	0.93	0.535	1.728

$m := 1 .. 5$

$G_m :=$

$\begin{bmatrix} L_{1,1} \\ L_{1,2} \\ L_{3,4} \\ L_{3,7} \\ L_{4,4} \end{bmatrix}$ $G = \begin{bmatrix} 0.8 \\ 0.4 \\ 0.326 \\ 0.419 \\ 0.368 \end{bmatrix}$

$n := 1 .. 4$

$H_n :=$

$\begin{bmatrix} R_{1,2} \\ R_{1,3} \\ R_{3,5} \\ R_{4,5} \end{bmatrix}$ $H = \begin{bmatrix} 1.5 \\ 2.15 \\ 1.744 \\ 1.974 \end{bmatrix}$

$p := 1 .. 11$ $q := 1 .. 11$

$A_{m,n} := \left(G_m\right)^2 + \left(H_n\right)^2 - 1$ $B_{m,n} := \left(G_m\right)^2 - \left(H_n\right)^2 + 1$

$Y_p :=$ $Y_q :=$

$$
\begin{array}{|c|}
\hline
\dfrac{1}{L_{5,5}} \\ \hline
\dfrac{1}{L_{5,6}} \\ \hline
\dfrac{1}{L_{6,5}} \\ \hline
\dfrac{1}{L_{6,6}} \\ \hline
\dfrac{1}{L_{7,5}} \\ \hline
\dfrac{1}{L_{8,5}} \\ \hline
\dfrac{1}{L_{9,6}} \\ \hline
\dfrac{1}{L_{10,5}} \\ \hline
\dfrac{1}{L_{3,4}} \\ \hline
\dfrac{1}{L_{4,4}} \\ \hline
\dfrac{1}{L_{5,4}} \\ \hline
\end{array}
\qquad
\begin{array}{|c|}
\hline
\dfrac{1}{L_{5,5}} \\ \hline
\dfrac{1}{L_{5,6}} \\ \hline
\dfrac{1}{L_{6,5}} \\ \hline
\dfrac{1}{L_{6,6}} \\ \hline
\dfrac{1}{L_{7,5}} \\ \hline
\dfrac{1}{L_{8,5}} \\ \hline
\dfrac{1}{L_{9,6}} \\ \hline
\dfrac{1}{L_{10,5}} \\ \hline
\dfrac{1}{L_{3,4}} \\ \hline
\dfrac{1}{L_{4,4}} \\ \hline
\dfrac{1}{L_{5,4}} \\ \hline
\end{array}
$$

$$
A = \begin{bmatrix}
1.89 & 4.262 & 2.682 & 3.535 \\
1.41 & 3.782 & 2.202 & 3.055 \\
1.356 & 3.729 & 2.148 & 3.001 \\
1.425 & 3.798 & 2.217 & 3.071 \\
1.386 & 3.758 & 2.178 & 3.031
\end{bmatrix}
\qquad
B = \begin{bmatrix}
-0.61 & -2.982 & -1.402 & -2.255 \\
-1.09 & -3.462 & -1.882 & -2.735 \\
-1.144 & -3.516 & -1.936 & -2.789 \\
-1.075 & -3.447 & -1.867 & -2.72 \\
-1.114 & -3.487 & -1.906 & -2.76
\end{bmatrix}
$$

$$L_{i,k} := \frac{l_k}{r_i}$$

$$
Y = \begin{array}{|c|c|}
\hline
 & 1 \\ \hline
1 & 3.125 \\ \hline
2 & 3.75 \\ \hline
3 & 1.583 \\ \hline
4 & 1.9 \\ \hline
5 & 2.208 \\ \hline
6 & 1.271 \\ \hline
7 & 4.925 \\ \hline
8 & 2.375 \\ \hline
9 & 3.071 \\ \hline
10 & 2.714 \\ \hline
11 & 5.357 \\ \hline
\end{array}
$$

$$
L = \begin{array}{|c|c|c|c|c|c|c|c|c|}
\hline
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline
1 & 0.8 & 0.4 & 0.2 & 0.7 & 1.2 & 1 & 0.9 & 3.3 \\ \hline
2 & 0.533 & 0.267 & 0.133 & 0.467 & 0.8 & 0.667 & 0.6 & 2.2 \\ \hline
3 & 0.372 & 0.186 & 0.093 & 0.326 & 0.558 & 0.465 & 0.419 & 1.535 \\ \hline
4 & 0.421 & 0.211 & 0.105 & 0.368 & 0.632 & 0.526 & 0.474 & 1.737 \\ \hline
5 & 0.213 & 0.107 & 0.053 & 0.187 & 0.32 & 0.267 & 0.24 & 0.88 \\ \hline
6 & 0.421 & 0.211 & 0.105 & 0.368 & 0.632 & 0.526 & 0.474 & 1.737 \\ \hline
7 & 0.302 & 0.151 & 0.075 & 0.264 & 0.453 & 0.377 & 0.34 & 1.245 \\ \hline
8 & 0.525 & 0.262 & 0.131 & 0.459 & 0.787 & 0.656 & 0.59 & 2.164 \\ \hline
9 & 0.162 & 0.081 & 0.041 & 0.142 & 0.244 & 0.203 & 0.183 & 0.67 \\ \hline
10 & 0.281 & 0.14 & 0.07 & 0.246 & 0.421 & 0.351 & 0.316 & 1.158 \\ \hline
\end{array}
$$

$$X_{p,q} := 1 + \frac{1 + \left(Y_q\right)^2}{\left(Y_p\right)^2}$$

$$F_{m,n} := \frac{1}{H_n} - \frac{1}{\pi \cdot H_n} \cdot \left[\operatorname{acos}\left(\frac{B_{m,n}}{A_{m,n}}\right) - \frac{1}{2 \cdot G_m} \cdot \left[\sqrt{\left(A_{m,n} + 2\right)^2 - \left(2 \cdot H_n\right)^2} \cdot \operatorname{acos}\left(\frac{B_{m,n}}{H_n \cdot A_{m,n}}\right) + B_{m,n} \cdot \operatorname{asin}\left(\frac{1}{H_n}\right) - \frac{\pi \cdot A_{m,n}}{2} \right] \right]$$

$$V_{p,q} := \frac{1}{2} \cdot \left[X_{p,q} - \sqrt{\left(X_{p,q}\right)^2 - 4 \cdot \left(\frac{Y_q}{Y_p}\right)^2} \right]$$

	1	2	3	4	5	6	7	8	9	10
1	2.102	2.542	1.359	1.472	1.602	1.268	3.586	1.68	2.068	1.857
2	1.766	2.071	1.249	1.328	1.418	1.186	2.796	1.472	1.742	1.595
3	5.294	7.008	2.399	2.839	3.344	2.043	11.074	3.649	5.162	4.338
4	3.982	5.172	1.971	2.277	2.628	1.724	7.996	2.84	3.89	3.318
5	3.208	4.089	1.719	1.945	2.205	1.536	6.179	2.362	3.139	2.716
6	7.666	10.327	3.171	3.854	4.639	2.619	16.638	5.112	7.46	6.181
7	1.444	1.621	1.145	1.19	1.242	1.108	2.041	1.274	1.43	1.345
8	2.909	3.67	1.622	1.817	2.042	1.464	5.477	2.177	2.85	2.483
9	2.141	2.597	1.372	1.489	1.623	1.277	3.677	1.704	2.106	1.887
10	2.461	3.044	1.476	1.626	1.798	1.355	4.428	1.901	2.416	2.136
11	1.375	1.525	1.122	1.161	1.205	1.091	1.88	1.231	1.364	1.292

X = (above)

$$F = \begin{bmatrix} 0.398 & 0.166 & 0.276 & 0.204 \\ 0.256 & 0.09 & 0.162 & 0.114 \\ 0.217 & 0.074 & 0.135 & 0.094 \\ 0.265 & 0.094 & 0.168 & 0.119 \\ 0.24 & 0.084 & 0.15 & 0.106 \end{bmatrix}$$

$F_{1,1} = 0.398 \qquad F_{1,2} = 0.166$

	1	2	3	4	5	6	7	8	9	10
1	0.727	0.852	0.227	0.321	0.424	0.148	0.938	0.482	0.712	0.6
2	0.591	0.767	0.164	0.235	0.314	0.106	0.919	0.361	0.575	0.463
3	0.883	0.922	0.537	0.661	0.75	0.39	0.956	0.786	0.879	0.84
4	0.869	0.915	0.459	0.594	0.701	0.318	0.954	0.747	0.863	0.816
5	0.849	0.906	0.385	0.519	0.638	0.259	0.951	0.693	0.842	0.781
6	0.893	0.926	0.605	0.711	0.783	0.464	0.958	0.812	0.889	0.857
7	0.378	0.533	0.099	0.142	0.191	0.064	0.817	0.221	0.365	0.287
8	0.835	0.9	0.349	0.478	0.599	0.233	0.95	0.658	0.827	0.756
9	0.737	0.857	0.233	0.33	0.435	0.152	0.939	0.494	0.723	0.613
10	0.796	0.883	0.286	0.4	0.517	0.188	0.945	0.579	0.785	0.693
11	0.324	0.46	0.084	0.121	0.163	0.054	0.744	0.188	0.313	0.245

V = (above)

$$F_{elfg} := \frac{1}{2 \cdot \left(l_1 + l_2\right)} \cdot \left(l_1 + l_2 + l_2 \cdot H_1 \cdot F_{2,1} - l_1 \cdot H_1 \cdot F_{1,1}\right)$$

$$F_{elfg} = 0.365$$

$$F_{elff} := .5 \cdot \left(H_1 \cdot F_{1,1} - H_2 \cdot F_{1,2}\right)$$

$$F_{elff} = 0.12$$

$$F_{fllb} := .5 \cdot H_3 \cdot \left(F_{4,3} - F_{3,3}\right)$$

$$F_{fllb} = 0.029$$

$$F_{flfb} := \frac{\left(r_5\right)^2}{\left(r_3\right)^2 - \left(r_4\right)^2} \cdot \left(V_{11,9} - V_{11,10}\right) - \frac{\left(r_4\right)^2}{\left(r_3\right)^2} \cdot \left(V_{9,10} - V_{11,10}\right)$$

$$F_{flfb} = 0.65$$

$$F_{fblb} := .5 \cdot \left(1 - H_4 \cdot F_{5,4}\right)$$

$$F_{fblb} = 0.396$$

$$F_{btp} := \frac{\left(r_7\right)^2}{\left(r_5\right)^2 - \left(r_6\right)^2} \cdot \left(V_{5,1} - V_{5,3}\right) - \frac{\left(r_8\right)^2}{\left(r_5\right)^2 - \left(r_6\right)^2} \cdot \left(V_{8,1} - V_{6,3}\right)$$

$$F_{btp} = 0.26$$

$$F_{bc} := \frac{\left(r_9\right)^2}{\left(r_5\right)^2 - \left(r_6\right)^2} \cdot \left(V_{7,2} - V_{7,4}\right) - \frac{\left(r_{10}\right)^2}{\left(r_5\right)^2 - \left(r_6\right)^2} \cdot \left(V_{8,2} - V_{8,4}\right)$$

$$F_{bc} = 0.579$$

this version has been modified for a converter with a 10 mil gap, and with contact resistances at both ends of the transition piece and a thermal path through the emitter sleeve and the bellows. The geometrical parameters are listed first:

$$\sigma := 5.67 \cdot 10^{-12} \cdot \frac{watt}{cm^2 \cdot K^4} \qquad \varepsilon_{ti} := 0.18 \qquad \varepsilon_{trim} := 0.2 \qquad \varepsilon_{cal} := 0.2 \qquad \varepsilon_{empara} := .2$$

$$d_{trans} := 5.08 \cdot cm \qquad d_{trim} := 6.9 \cdot cm \qquad l_{trans} := 1.905 \cdot cm$$

$$\varepsilon_b := .2$$

$$A_{ee} := 1 \cdot cm^2 \qquad A_{trans1} := 5.31 \cdot cm^2$$

$$A_{emedge} := 2 \cdot \pi \cdot r_1 \cdot (l_1 + l_2) \qquad A_{emedge1} := 2 \cdot \pi \cdot r_1 \cdot l_1 \qquad A_{backer} := \pi \cdot (r_1)^2 \qquad A_{er} := \pi \cdot (r_1)^2$$

$$A_{emedge} = 7.54 \cdot cm^2 \qquad A_{emedge1} = 5.027 \cdot cm^2 \qquad A_{backer} = 3.142 \cdot cm^2 \qquad A_{er} = 3.142 \cdot cm^2$$

$$A_{flange} := 20.1 \cdot cm^2 \qquad A_{iflange} := 2 \cdot \pi \cdot r_4 \cdot (l_3 + l_4 - l_2) \qquad A_{flipl} := 2 \cdot \pi \cdot r_3 \cdot l_3 \qquad A_{fbl} := 2 \cdot \pi \cdot r_4 \cdot l_4$$

$$A_{flange} \qquad A_{iflange} = 5.969 \cdot cm^2$$

$$A_{flipl} = 2.702 \cdot cm^2 \qquad A_{fbl} = 8.357 \cdot cm^2$$

$$A_{bel} := \pi \cdot \left[(r_5)^2 - (r_4)^2 \right] \qquad A_{cone} := 1.41 \cdot \pi \cdot \left[(r_2)^2 - (r_1)^2 \right] \qquad A_{flipf} := \pi \cdot \left[(r_3)^2 - (r_4)^2 \right]$$

$$A_{bel} = 32.837 \cdot cm^2 \qquad A_{cone} = 5.537 \cdot cm^2 \qquad A_{flipf} = 3.181 \cdot cm^2 \qquad A_{fbl} := 2 \cdot \pi \cdot r_4 \cdot l_4$$

$$A_{tp} := \pi \cdot \left[(r_7)^2 - (r_8)^2 \right]$$

$$A_{tp} = 14.756 \cdot cm^2$$

$$A_{call} := 2 \cdot \pi \cdot r_9 \cdot l_8 \qquad A_{calf} := \pi \cdot \left[(r_9)^2 - (r_{10})^2 \right] \qquad A_{cal} := A_{call} + A_{calf} \qquad A_{fbl} = 8.357 \cdot cm^2$$

$$A_{call} = 102.117 \cdot cm^2 \qquad A_{calf} = 50.684 \cdot cm^2 \qquad A_{cal} = 152.801 \cdot cm^2$$

the following are the calculated thermal resistances:

$$A_{colrad} := 5.1 \cdot cm^2$$

$$N := 0 \qquad R_s := 8.1 \cdot \frac{K}{watt} \qquad R_b := 23 \cdot \frac{K}{watt} \qquad k_{interface} := .17 \cdot \frac{watt}{cm^2 \cdot K} \qquad A_{trans2} := 8.04 \cdot cm^2$$

n and b for electron cooling

o and c for TI electric power density

$$n := .0423 \cdot \frac{watt}{K \cdot cm^2} \qquad b := 58.6 \cdot \frac{watt}{cm^2}$$

$$o := 0.0135 \cdot \frac{watt}{K \cdot cm^2} \qquad c := 20.46 \cdot \frac{watt}{cm^2}$$

p and d for AMTEC parasitic loss   q and e for AMTEC elec. pwr

r and f for AMTEC current
density

$$q := 1.2 \cdot 10^{-2} \cdot \frac{watt}{K}$$

$$p := 5 \cdot 10^{-2} \cdot \frac{watt}{K} \qquad d := 34.3 \cdot watt$$

$$e := 8.9 \cdot watt$$

$$r := 2 \cdot 10^{-2} \cdot \frac{amp}{K}$$

$$G := 1 \cdot \frac{w_\cdot}{an}$$

$$T_{res} := 573 \cdot K$$

$$f := 10 \cdot amp$$

these are starting points for the calculations:

$$T_e := 1800 \cdot K \qquad T_c := 1100 \cdot K \qquad T_a := 1000 \cdot K \quad T_b := 1000 \cdot K \qquad T_1 := 400 \cdot K \qquad T_{cal} := 400 \cdot K$$

$$T_{amb} := 325 \cdot K \qquad T_f := 1200 \cdot K \qquad T_{bel} := 1000 \cdot K \qquad T_{trim} := 500 \cdot K$$

these are the 'input parameters':

$$R_{res} := 100 \cdot \frac{K}{watt}$$

$$Q_{input} := 318 \cdot watt \qquad Q_{intrim} := 126 \cdot watt$$

$$R_{flange} := 100 \cdot \frac{K}{watt} \qquad R_1 := 1 \cdot \frac{K}{watt}$$

this is the solve block containing the nine nodal equations in nine unknown temperatures:

given

at emitter node:

$$Q_{input} = \sigma \cdot \varepsilon_{empara} \cdot \left[A_{backer} \cdot \frac{1}{1+N} + A_{emedge} \cdot (1 - F_{elff}) - A_{emedge1} \cdot F_{elfg} \right] \cdot \left(T_e^4 - T_{amb}^4 \right) \cdots$$

$$+ (o \cdot T_e - c) \cdot A_{ee} + \sigma \cdot \varepsilon_{ti} \cdot A_{er} \cdot \left(T_e^4 - T_c^4 \right) + (n \cdot T_e - b) \cdot A_{ee} + \frac{T_e - T_f}{R_s} \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot \left(A_{emedge} \cdot F_{elfg} + A_{emedge1} \cdot F_{elff} \right) \cdot \left(T_e^4 - T_f^4 \right)$$

at flange node:

$$\frac{T_e - T_f}{R_s} \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot \begin{pmatrix} A_{emedge} \cdot F_{elfg} \cdots \\ + A_{emedge1} \cdot F_{elff} \end{pmatrix} \cdot \left(T_e^4 - T_f^4 \right)$$

$$= \frac{T_f - T_{bel}}{R_b \cdot 0.5} \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot \left[\begin{array}{l} A_{cone} \cdot \left(1 - \dfrac{A_{emedge}}{A_{cone}} \cdot F_{elfg} \right) \cdots \\[2ex] + \pi \cdot \left(r_3^2 - r_2^2 \right) \cdot \left(1 - \dfrac{2 \cdot r_1 \cdot l_1}{r_3^2 - r_2^2} \cdot F_{elff} \right) \cdots \\[2ex] + 2 \cdot \pi \cdot r_3 \cdot l_3 \cdot \left(1 - F_{fllb} \right) \cdots \\[1ex] + \pi \cdot \left(r_3^2 - r_4^2 \right) \cdot \left(1 - F_{flfb} \right) \cdots \\[1ex] + 2 \cdot \pi \cdot r_4 \cdot l_4 \cdot \left(1 - F_{fblb} \right) \end{array} \right] \cdot \left(T_f^4 - T_{amb}^4 \right)$$

$$+ \frac{T_f - T_{amb}}{R_{flange}} + \frac{T_f - T_{res}}{R_{res}} \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot A_{iflange} \cdot \left(T_f^4 - T_c^4 \right) \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot \left[\begin{array}{l} A_{flipl} \cdot F_{fllb} \cdots \\ + A_{flipf} \cdot F_{flfb} \cdots \\ + A_{fbl} \cdot F_{fblb} \end{array} \right] \cdot \left(T_f^4 - T_{bel}^4 \right)$$

at bellows node:

$$\frac{T_f - T_{bel}}{R_b \cdot 0.5} \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot \begin{bmatrix} A_{flipl} \cdot F_{fllb} \cdots \\ + A_{flipf} \cdot F_{flfb} \cdots \\ + A_{fbl} \cdot F_{fblb} \end{bmatrix} \cdot \left(T_f^4 - T_{bel}^4 \right) \cdots$$

$$+ \frac{T_c - T_{bel}}{R_b \cdot 0.5} + \sigma \cdot \varepsilon_{trim} \cdot A_{tp} \cdot F_{btp} \cdot \left(T_b^4 - T_{bel}^4 \right)$$

$$= \sigma \cdot \varepsilon_{empara} \cdot \left[A_{bel} \cdot \left(2 - F_{btp} - F_{bc} \right) \cdots \right] \cdot \left(T \right.$$

$$+ \left[- \begin{bmatrix} A_{flipl} \cdot F_{fllb} \cdots \\ + A_{flipf} \cdot F_{flfb} \cdots \\ + A_{fbl} \cdot F_{fblb} \end{bmatrix} \right]$$

$$+ \sigma \cdot \varepsilon_{cal} \cdot A_{bel} \cdot F_{bc} \cdot \left(T_{bel}^4 - T_{cal}^4 \right)$$

at collector node:

$$\begin{bmatrix} \sigma \cdot \varepsilon_{ti} \cdot A_{er} \cdot \left(T_e^4 - T_c^4 \right) + \left(n \cdot T_e - b \right) \cdot A_{ee} \cdots \\ + \sigma \cdot \varepsilon_{empara} \cdot A_{iflange} \cdot \left(T_f^4 - T_c^4 \right) \end{bmatrix} = \frac{T_c - T_{bel}}{R_b \cdot 0.5} + A_{trans1} \cdot k_{interface} \cdot \left(T_c - T_b \right) \cdots$$

$$+ \sigma \cdot \varepsilon_{empara} \cdot A_{colrad} \cdot \left(T_c^4 - T_{amb}^4 \right)$$

at middle of transition piece:

$$A_{trans1} \cdot k_{interface} \cdot \left(T_c - T_b \right) = \frac{A_{trans2} \cdot k_{interface}}{1} \cdot \left(T_b - T_a \right) \cdots$$

$$+ \sigma \cdot \varepsilon_{trim} \cdot \pi \cdot d_{trans} \cdot l_{trans} \cdot \left(T_b^4 - T_{trim}^4 \right) \cdots$$

$$+ \sigma \cdot \varepsilon_b \cdot A_{bel} \cdot F_{btp} \cdot \left(T_b^4 - T_{bel}^4 \right)$$

at trim collar node:

$$\sigma \cdot \varepsilon_{trim} \cdot \pi \cdot d_{trans} \cdot l_{trans} \cdot \left(T_b^4 - T_{trim}^4 \right) + Q_{intrim} = \sigma \cdot \varepsilon_{cal} \cdot \pi \cdot d_{trim} \cdot l_{trans} \cdot \left(T_{trim}^4 - T_{cal}^4 \right)$$

at AMTEC node:

$$\frac{A_{trans2} \cdot k_{interface}}{1} \cdot \left(T_b - T_a \right) = \left(r \cdot T_a - f \right) \cdot G + \left(p \cdot T_a - d \right) + \left(q \cdot T_a - e \right)$$

at condensor node:

$$\left(r \cdot T_a - f \right) \cdot G + \left(p \cdot T_a - d \right) = \frac{T_1 - T_{amb}}{R_1}$$

at calomiter node:

$$\varepsilon_{cal} \cdot A_{bel} \cdot F_{bc} \cdot \left(T_{bel}^4 - T_{cal}^4 \right) \cdots$$

$$+ \varepsilon_{cal} \cdot \pi \cdot d_{trim} \cdot l_{trans} \cdot \left(T_{trim}^4 - T_{cal}^4 \right)$$

$$= \varepsilon_{cal} \cdot \left(A_{cal} - A_{bel} \cdot F_{bc} \right) \cdot \left(T_{cal}^4 - T_{amb}^4 \right)$$

$$
\begin{bmatrix}
T_e \\
T_c \\
T_f \\
T_{bel} \\
T_b \\
T_a \\
T_{trim} \\
T_1 \\
T_{cal}
\end{bmatrix}
:= \text{Minerr}\left(T_e, T_c, T_f, T_{bel}, T_b, T_a, T_{trim}, T_1, T_{cal}\right)
\qquad
\begin{bmatrix}
T_e \\
T_c \\
T_f \\
T_{bel} \\
T_b \\
T_a \\
T_{trim} \\
T_1 \\
T_{cal}
\end{bmatrix}
=
\begin{bmatrix}
1.95 \cdot 10^3 \\
1.268 \cdot 10^3 \\
1.327 \cdot 10^3 \\
1.06 \cdot 10^3 \\
1.232 \cdot 10^3 \\
1.199 \cdot 10^3 \\
1.312 \cdot 10^3 \\
364.641 \\
934.451
\end{bmatrix} \cdot K
$$

$$ERR = 0.004$$

$$Q_{colrad} := \sigma \cdot \varepsilon_{empara} \cdot A_{colrad} \cdot \left(T_c^4 - T_{amb}^4\right)$$
$$Q_{colrad} = 14.891 \cdot watt$$

$$Q_{bellpara} := \sigma \cdot \varepsilon_{empara} \cdot \left[A_{bel} \cdot \left(2 - F_{btp} - F_{bc}\right) \cdots + \left[-\left[A_{flipl} \cdot F_{fllb} \cdots + A_{flipf} \cdot F_{flfb} \cdots + A_{fbl} \cdot F_{fblb} \right] \right] \right] \cdot \left(T_{bel}^4 - T_{amb}^4\right)$$

$$Q_{bellpara} = 46.382 \cdot watt$$

$$Q_{flngpara} := \sigma \cdot \varepsilon_{empara} \cdot \left[A_{cone} \cdot \left(1 - \frac{A_{emedge}}{A_{cone}} \cdot F_{elfg}\right) \cdots + \pi \cdot \left(r_3^2 - r_2^2\right) \cdot \left(1 - \frac{2 \cdot r_1 \cdot l_1}{r_3^2 - r_2^2} \cdot F_{elff}\right) \cdots + 2 \cdot \pi \cdot r_3 \cdot l_3 \cdot \left(1 - F_{fllb}\right) \cdots + \pi \cdot \left(r_3^2 - r_4^2\right) \cdot \left(1 - F_{flfb}\right) \cdots + 2 \cdot \pi \cdot r_4 \cdot l_4 \cdot \left(1 - F_{fblb}\right) \right] \cdot \left(T_f^4 - T_{amb}^4\right)$$

$$Q_{flngpara} = 64.516 \cdot watt$$

$$Q_{flangcond} := \frac{T_f - T_{amb}}{R_{flange}}$$

$$Q_{flangcond} = 10.019 \cdot watt$$

$$Q_{emitterbacklosses} := \sigma \cdot \varepsilon_{empara} \cdot \left[A_{backer} \cdot \frac{1}{1+N} + A_{emedge} \cdot (1 - F_{elff}) - A_{emedge1} \cdot F_{elfg} \right.$$

$$Q_{emitterbacklosses} = 130.106 \cdot watt$$

$$Q_{electric} := (o \cdot T_e - c) \cdot A_{ee} \qquad Q_{electron} := (n \cdot T_e - b) \cdot A_{ee}$$

$$Q_{electric} = 5.863 \cdot watt \qquad Q_{electron} = 23.878 \cdot watt$$

$$Q_{trim} := \sigma \cdot \varepsilon_{trim} \cdot \pi \cdot d_{trans} \cdot l_{trans} \cdot \left(T_b^4 - T_{trim}^4 \right)$$

$$Q_{trim} = -22.797 \cdot watt$$

$$Q_{cal} := \sigma \cdot \varepsilon_{cal} \cdot \pi \cdot d_{trim} \cdot l_{trans} \cdot \left(T_{trim}^4 - T_{amb}^4 \right)$$

$$Q_{cal} = 138.386 \cdot watt$$

$$Q_{calnew} := \sigma \cdot \varepsilon_{cal} \cdot \left(A_{cal} - A_{bel} \cdot F_{bc} \right) \cdot \left(T_{cal}^4 - \right.$$

$$Q_{rade} := \sigma \cdot \varepsilon_{ti} \cdot A_{er} \cdot \left(T_e^4 - T_c^4 \right)$$

$$Q_{calnew} = 1.14 \cdot 10^9 \cdot gm \cdot cm^2 \cdot sec^{-3}$$

$$Q_{rade} = 38.054 \cdot watt$$

$$or := 114 \cdot watt$$

$$Q_{amtecelec} := (q \cdot T_a - e) \qquad Q_{amtecpara} := (p \cdot T_a - d)$$

$$Q_{amtecthru} := (r \cdot T_a - f) \cdot G$$

$$Q_{amtecelec} = 5.49 \cdot watt$$

$$Q_{amtecpara} = 25.658 \cdot watt \qquad Q_{amtecthru} = 13.983 \cdot watt$$

heat balance:

$$\left(Q_{input} + Q_{intrim} \right) - Q_{flngpara} - Q_{bellpara} - Q_{cal} - Q_{amtecthru} - Q_{amtecelec} \cdots = -11.293 \cdot w_{\ell}$$
$$+ \left(-Q_{amtecpara} - Q_{electric} - Q_{emitterbacklosses} - Q_{colrad} - Q_{flangcond} \right)$$

$$\eta_{hybrid} := \frac{Q_{electric} + Q_{amtecelec}}{Q_{electron} + Q_{rade} - Q_{trim}}$$

$$\eta_{hybrid} = 13.399 \cdot \%$$

Conclusion

The task of mathematical modeling of the cascade consisting of TIEC and AMTEC has been complete. This is, however, not the end of the research for modeling the cascade. The data from the new cascade device according to the present model has not been yet obtained (at the time of preparing this report). The adjustments of certain parameters and re-examining some terms in the light of the data is expected to be performed. This model is, however, the initial but very essential step of the overall research on AMTEC TIEC cascade development. Further data taking may suggest quite likely that more detail modeling of the cascade is needed.

It may be appropriate to make a remark here that while AMTEC by itself could be used for general solar thermal equipment and device its cascading with TIEC enhances its role in improving the efficiency of the system as a whole in a global manner. It is thus obvious from these studies that furthering the future research and development of the cascade are warranted both experimentally and theoretically.
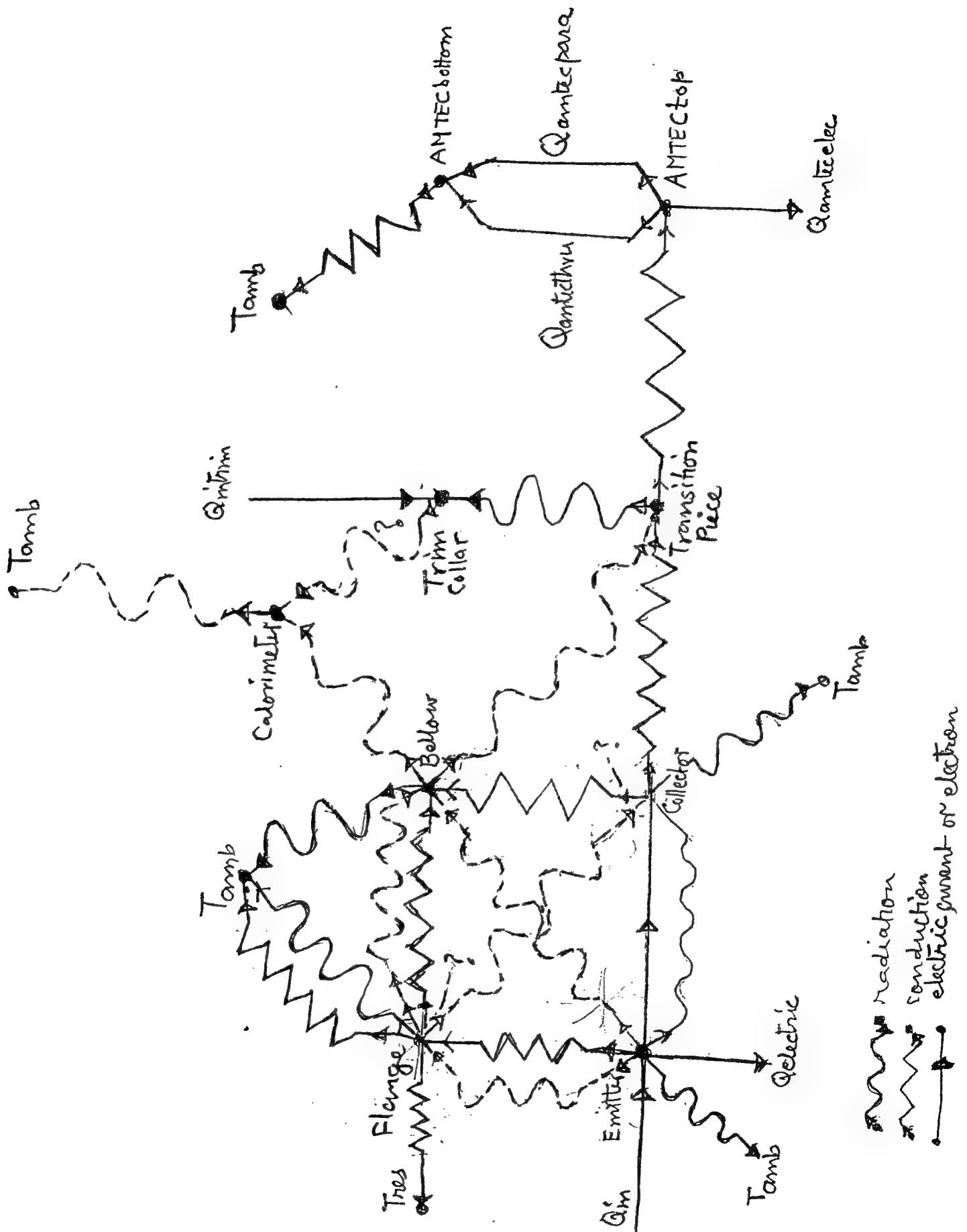
Acknowledgment

STATIC ENERGY CONVERSION FOR SPACE POWER

NO	DEVICE	HEAT SOURCE	TEMP (K)	WORKING FLUID	POWER (W)	SPECIFIC POWER (We/kg)	POWER DENSITY (W/cm²)	UNIT COST ($/W)	LIFE TIME (Y)	EFFICIENC Y %	TECHNOLOGY STATUS
1	TIEC		1970	Cs vapor			8		15,2	17	
2	TOEC	RTGS	1980 2	NONE (SiGe)					>7 10 MAX	7	Voyager Flight, Reliable Well-Proven
3	TPV	Solar Conc, Nuclear RTG GPHS	1470 2200 2400	NONE		6-10	5			7-14 (30)	Exploratory State
4	AMTEC	Solar, Nuclear	1400	Na, K Vapor & Liquid	W-MW	13-16 20(Proj) for <1kw	5	1000	10	19 13(expt) 20(th)	Most Advanced Stage Leading to Space Flight Test
5	HYTEC	Solar, Nuclear	900	Li, LiNa Liquid H		Light Wt			20 (Proj)		Losses at Power Producing Step are yet to be Determined
6	TAPC	Solar, Combustion	1100	Gas, Li, He Na Li	1kw					15-21	Early State of Development
7	LMMHD	Reactor, RTG	1123 - 1673	mixture of Liq & Vapor	100 kw - 100 MW					10-15 (th) <10(ex)	Exploratory Stage

Table 1

Emitter, W

Emitter sleeve, Mo/Re

Support rod, Mo

T_e

Emitter flange, Mo

T_f

T_{bel} Bellows, Ni

Insulator, alumina

T_c

Transition Piece, Ni

Trim collar, Ni

Collector, Ni

Calorimeter collar, Ni

T_b

T_{trim}

T=300 K

tact
stances

T_a

Insulator, boron nitride

AMTEC heat pipe

T_1

AMTEC condenser

T=300 K
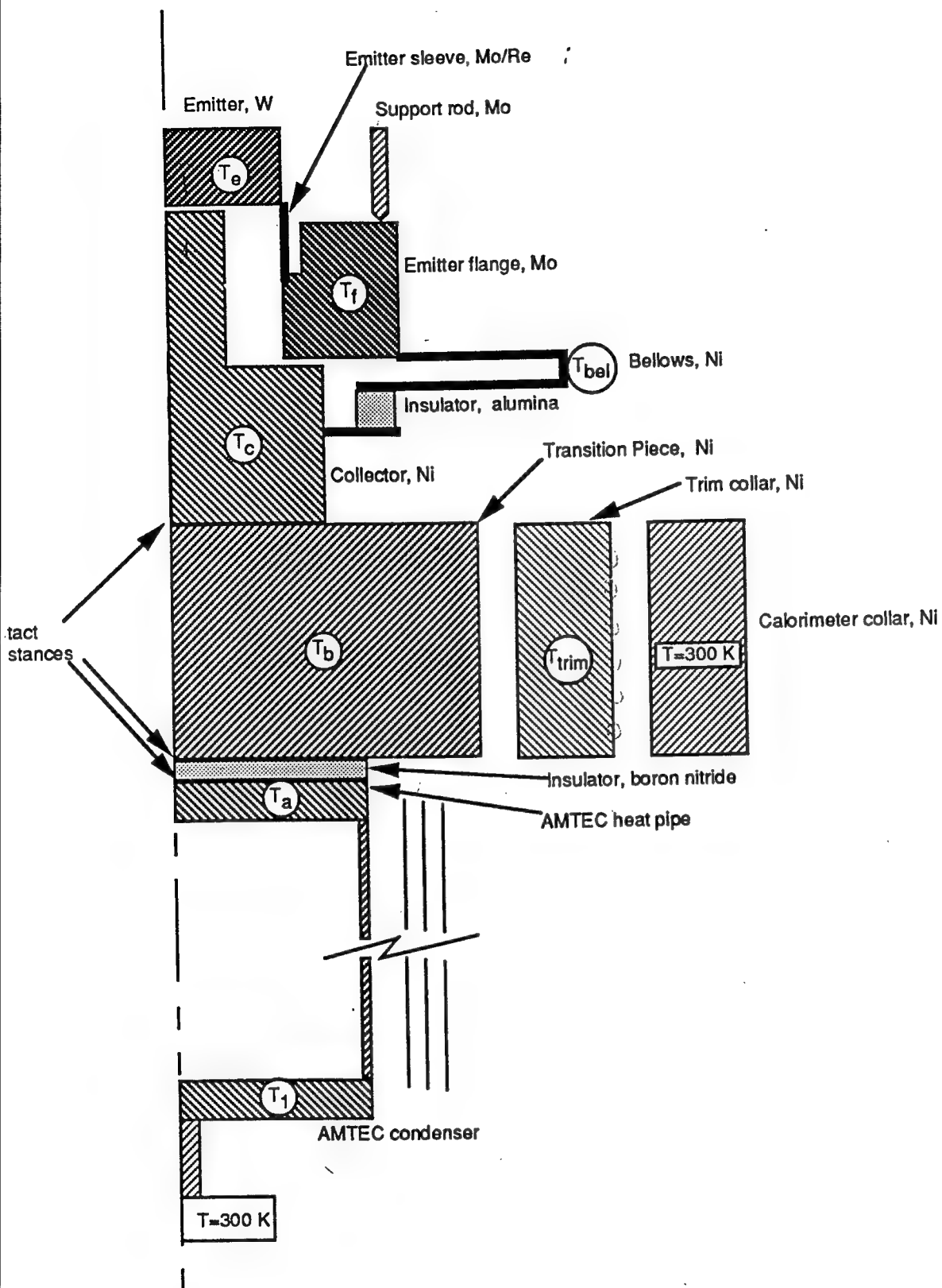
SCHEMATIC OF THERMAL MODEL OF CASCADE TEC/AMTEC CONVERTER

FIG 1

Labels visible in figure: AMTEC bottom, Q antec para, AMTEC top, Q antic elec, Tamb, Q antic thru, Q'in Trim, Tamb, Trim Collar, Transition Piece, Calorimeter, Bellows, Tamb, Collector, Tamb, Flange, Emitter, Q electric, Tres, Q'in, Tamb

Legend: radiation, conduction, electric current or electron

23-18

FIG 2

ESTIMATING THE SPATIAL DENSITY AND COLLISION RATES FOR THE NEAR-EARTH MAN-MADE ORBITAL DEBRIS ENVIRONMENT

Ronald A. Madler
Assistant Professor
Department of Aerospace Engineering

Embry-Riddle Aeronautical University
3200 Willow Creek Road
Prescott, Arizona 86301-3720

Final Report for:
Summer Faculty Research Program
Phillips Laboratory

August 1995

ESTIMATING THE SPATIAL DENSITY AND COLLISION RATES FOR THE NEAR-EARTH MAN-MADE ORBITAL DEBRIS ENVIRONMENT

Ronald A. Madler
Assistant Professor
Department of Aerospace Engineering
Embry-Riddle Aeronautical University

Abstract

Determining the current debris population is the first step toward quantifying the hazard posed by natural and man-made debris. Ground-based sensors currently track and maintain orbits of the largest objects in this population, while statistical estimates of the population of smaller objects have also recently been made. The second step toward quantifying the hazard is to use models to estimate future debris populations and the effects they could have on future space operations. The USAF Phillips Laboratory has developed and is improving such a model as part of its space debris research program. The model is called the Debris Environment and Effects Program (DEEP). DEEP is being developed to: 1) estimate a range of possible future debris environments based on a variety of parameters, 2) determine the collision hazard for specific vehicles operating in these environments, 3) determine the effects of potential collisions on vehicle mission performance and 4) evaluate the effects of implementing various options for mitigating man-made debris.

An essential part of any debris environment model is the method used to determine the collision rates between man-made space objects. The methodology behind and some results of the spatial density algorithms used within the DEEP model are discussed.

ESTIMATING THE SPATIAL DENSITY AND COLLISION RATES FOR THE NEAR-EARTH MAN-MADE ORBITAL DEBRIS ENVIRONMENT

Ronald A. Madler

Introduction

There exists a need to quantify and characterize the hazard posed to orbiting assets by orbital space debris. There are currently over 7000 objects being tracked by the Department of Defense (DoD) Space Surveillance Network (SSN), and a much larger population of objects below the detection threshold is known to exist.[1] The USAF Phillips Laboratory has been developing the Debris Environment and Effects Program (DEEP) to estimate the debris environment and the hazards to LEO space operations, for both the present and future.[2,3] Determining the spatial density and the collision rates between objects in space is a crucial part of the process to determine the hazard to LEO operations. It is essential to have accurate estimates of the collision rates between orbiting objects so that simulations of the future space environment are as realistic as possible. This paper briefly introduces the models to determine the spatial density and collision rates between space resident objects.

Discussion of Problem

DEEP estimates the current and future debris environments by modeling known and anticipated debris sources and sinks. Future unintentional collisions are expected to be a large source of future space debris.[4,5,6] In order for these unintentional collisions to be modeled well, reasonable estimates of the collision rates among all sized objects must be made. DEEP models the orbital environment by maintaining orbital elements and associated information for discrete objects representing intact objects (IO) and fragments from debris clouds (DC).[7,8] These orbital elements allow one to calculate an average spatial density for the near-Earth environment at a certain time. Using the relative velocities between the orbiting objects, which is a function of the inclination of the interacting objects, one can calculate an expected collision rate.[9]

Since collision events have played a major role in forming our solar system, planetary scientists have been interested in calculating the collision rates between solar system objects for some time.[10,11,12,13] Two methods for determining collision rates are generally used which assume that the periapse and ascending node are randomly oriented. The first method, developed by

Opik and later generalized by Wetherill and Greenberg, solves for the fraction of time that the two objects are within a small volume of space such that they overlap with one another. The implementation of this method turns out to be computationally intensive, yet it has been used for many planetary studies.[14,15,16,17] Similar methods were used in early orbital debris studies to determine the collision rate between Earth orbiting objects.[18,19]

The other basic approach is analogous to methods used to predict collision rates in the kinetic theory of gases. This method assumes that the objects are randomly located within a volume of space and examines the area swept out by the motion of the objects. Kessler uses the spatial density method to derive equations equivalent to the first method, yet which yields a much quicker solution.[10] This basic method is to find the spatial density of the orbiting objects as a function of radius and inertial latitude, and then use the approximate relative velocity between orbiting objects to obtain a collision rate.

The collision rate model for DEEP uses a method more closely related to that of Kessler, yet does not require that the ascending node (Ω) and the argument of periapse (ω) be randomly oriented. While the assumption of randomness is valid under many circumstances, the DEEP model includes the ability to analyze the debris environment for time scales short enough for this assumption to no longer be accurate. With reasonable estimates for the ascending node (Ω) and the argument of periapse (ω) one can find significant variation within the spatial density and resulting flux seen by a spacecraft.

The inclusion of ascending node and periapse information allows one to develop the spatial density with more fidelity in terms of the radius, as well as the inertial latitude and longitude. While the three-dimensional volumetric grid including a right ascension term generally will not be used to calculate the collision rates for long-term modeling due to the precession of the ascending node and periapse within the longer time steps of long-term models, it will be used to examine short- to medium-term hazards to space assets. The three dimensional gridding allows for a better understanding of the actual velocity and angular distributions of possible debris impacts for a specific mission.

Estimating Collisions

Unintentional collisions are expected to be a major source of debris in the next century. [4,5,6] Collision rates are estimated by determining the spatial density within a gridded volume of near-Earth space and calculating a flux using the appropriate relative velocity distribution for each spatial volume. The number of expected collisions during a time interval is calculated using the

spatial density of the two object populations, the mean physical size of the objects, and the relative velocity between their orbits. This expected collision rate between two populations is given in Equation (1), where: N is the expected number of impacts; σ is the collisional cross section; S_1 and S_2 are the spatial densities of the populations; V_{rel} is the relative impact velocity; and ΔVolume is the volume of the spatial density element one is interested in.

$$N/t = \sum (\sigma * S_1 * S_2 * V_{rel} * \Delta(\text{Volume})) \qquad (1)$$

Once the expected collision rate has been found, the actual probability for a certain number of collisions occurring during a given time interval can be calculated with Poisson or binomial distribution statistics, depending upon the value of the collision rate.

An accurate estimate for each of the terms in Eqn. (1) is necessary in order to estimate the collision rate. The collisional cross section depends upon the shapes and sizes of the interacting objects. Since many objects have irregular shapes, most researchers has assumed that all objects can be treated as being roughly spherical. This has been shown to underestimate the cross sectional area, and hence the collision rate.[9]

The relative velocity between orbiting objects is a function of the interaction geometry and thus the orbital parameters. For the present trackable orbital population the *average* relative impact velocity in low Earth orbit (LEO) is about 10 km/s, but for the interaction of individual objects obviously depends on the orbit geometry.[9] Since the impact velocity depends upon the interaction geometry, it will vary with latitude. This is taken into account by determining the velocity distribution as a function of latitude.[20] Work to define the velocity distribution as a function of latitude is planned as a part of a future AFOSR proposal.

The volume term in Eqn. (1) is easily determined for a region in space. This leaves the spatial density terms. The next section descries a method to determine the spatial density when the orbital parameters are all known.

Spatial Density Calculation

This section describes the methods for determining the spatial density of objects in Earth orbit. The spatial density calculations will be computed such that the spatial density can be a function of: altitude, latitude, and/or right ascension. The true anomaly (f, v, or θ depending on

convention) is assumed to be random within the orbit plane for each method since this is an extremely fast changing variable and is subject to many perturbing forces. Other approximations for calculating average spatial densities will be examined later in this document.

The method for spatial density calculations one should use depends greatly upon the purpose for which the calculations are to be made. When looking at collision rates over long periods of time, the spatial density will be computed as a function of radius and latitude only, due to perturbing forces. However, over short time scales there will be significant variations in the spatial density as a function of longitude also. This is exacerbated when taking debris from on-orbit fragmentations into account, which produce a large number of objects. These debris "clouds" will disperse over time, but can produce significant spatial density variations in the meantime.

The first described spatial density technique is similar to a Teledyne Brown method when the spatial density is a function of all three parameters (altitude, latitude and right ascension).[21] The spatial density calculated with this method depends upon the semimajor axis (a), eccentricity (e), inclination (i), longitude of the ascending node (Ω), and the argument of periapse (ω) for the objects in orbit. Due to the oblateness of the Earth (J2), the latter two terms will experience significant perturbations, hence the implementation of these specific routines are only meant to provide short- to medium-term spatial densities for the on-orbit population. For any long-term collision probabilities, this first technique should NOT be used without step sizes short enough so that precession does not change the ascending node and periapse distributions significantly.

Spatial Density Calculations With Known Orbital Parameters

If the longitude of the ascending node (Ω) and the argument of periapse (ω) are know for each orbiting object, an average spatial density matrix can be set up. This is most easily done by starting from the perigee point and calculating the elapsed transit time along the orbit to each volume element boundary. The boundary crossing times for each spatial parameter (i.e. radius, latitude, and right ascension (sometimes called a longitude in this document)) are calculated separately. By taking the three time and position arrays generated and stepping from the perigee point all the way around the orbit back to the perigee point the appropriate time, and hence spatial density, can be recorded for each volume element. This approach gives accurate results and is summarized below:

1) Establish an array of transit times from perigee to all inclusive *radius (altitude)* boundaries. This is accomplished by determining the eccentric anomaly angle, E, using the expression

$$E = \cos^{-1}\left(\frac{a - R}{a e}\right) \tag{2}$$

where R is the radius of interest. This expression can be derived from the polar form of the conic equation and the geometry of the ellipse, however, the value for the eccentric anomaly will always be returned with values between 0 and π on any computer. Thus, care must be exercised when determining the eccentric anomaly and transit time, but it can be noted that motion is symmetric about the major axis. The transit time is then given by

$$\Delta t = \frac{1}{n}(E - e \sin E) \tag{3}$$

where n is the mean motion of the orbit. When the eccentric or true anomaly is greater than π and less than 2π (i.e. $180° \le f \le 360°$), the Δt value must be properly adjusted by subtracting Δt from the orbital period.

2) Establish an array of transit times from perigee to all inclusive *latitude* boundaries. This is accomplished by first determining the true anomaly (*f*) at the latitude boundary, which is determined using the argument of latitude (u) where u = ω + f, such that

$$\sin(u) = \frac{\sin \beta}{\sin i} \, . \tag{4}$$

β is the latitude and i is the inclination. Since the value for the argument of latitude will be returned between $-\pi/2$ and $\pi/2$ (-90° and 90°), care must be exercised to find the correct quadrant for the true and eccentric anomalies. Another equivalent expression for the true anomaly follows.

$$f(\beta) = \sin^{-1}\left(\frac{\sin \beta}{\sin i}\right) - \omega \tag{5}$$

where i is the inclination and ω is the argument of perigee. To obtain a value in the range $\pi \le f \le 2\pi$ (0° $\le f \le$ 360°), it may be necessary to subtract f from π(180°) due to trigonometric quadrant considerations.

The eccentric anomaly is then calculated unambiguously using

$$E = 2\tan^{-1}\left(\sqrt{\frac{1-e}{1+e}}\,\tan\left(\frac{f}{2}\right)\right)$$
(6)

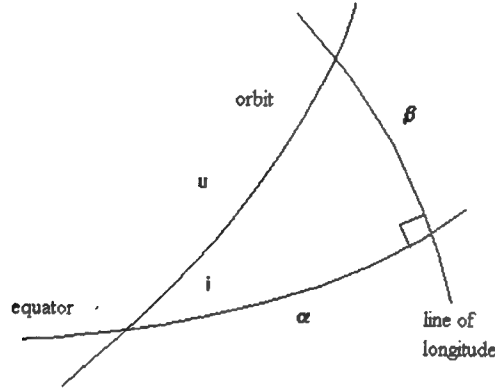The transit time to the latitude boundary is then calculated using Eqn. (3).

Figure 1 - Spherical geometry of the latitude encounter

3) Establish an array of transit times from perigee to all inclusive *right ascension* boundaries. This is accomplished by determining the true anomaly (f) at the right ascension boundary $\phi = (\Omega + \alpha)$ according to the general form

$$f(\phi) = \tan^{-1}\left(\frac{\tan(\phi - \Omega)}{\cos i}\right) - \omega$$
(7)

where Ω is the right ascension of the ascending node and is related to the figure by $\alpha = \phi - \Omega$.. A series of trigonometric quadrant conditions (which depend on the values of ω and Ω) must be checked to properly place f within the range $0° \le f \le 360°$. The eccentric anomaly is then calculated using Eqn. (6) whose result then feeds into the transit time expression given by Eqn. (3).

4) Starting from the perigee point, find the minimum of the transit times among the first three spatial parameter boundaries encountered. The minimum represents the *next*

boundary encountered regardless of whether it is a radius, latitude, or right ascension boundary. Calculate the spatial density (D) for this specific volume element with known transit time (Δt) using the general expression

$$D = S \cdot T \cdot N \cdot \left(\frac{\Delta t}{P \cdot V} \right) \tag{8}$$

P is the orbital period, V is the volume of the volume element under consideration, and N is the number of intact object orbits sharing the same orbital parameters ($N = 1$ for all breakup fragment orbits and for most all intact object orbits). S is the scaling factor to account for the density adjustments resulting from the orbit being a "trace" orbit if it is a fragmentation debris fragment ($S = 1$ if it is an intact launched object or operational debris), and T is the percentage of the time step the orbit has contributed to the environment ($T = 1$ if the object was resident in the environment at or before the previous time step). The expression for the volume is easily derived from the differential volume element given by

$$dV = R^2 \cos\beta \, dR \, d\beta \, d\phi \tag{9}$$

$$V = \frac{\Delta\phi}{3} \left[\left((R + \Delta R)^3 - R^3 \right) \left(\sin(\beta + \Delta\beta) - \sin(\beta) \right) \right] \tag{10}$$

5) Repeat step (4) above until the perigee point is again reached, completing the cycle around the orbit.

6) Repeat step (1)-(5) for all discrete orbits in the environment. The resulting spatial density (SPD) array will then represent a mapping of the density from all orbits found in the LEO environment.

This spatial density algorithm has been incorporated into a FORTRAN 90 subroutine called SD. The following discussion explains how to determine the boundary crossing times needed in the above method.

1) Altitude (radius) boundaries - An array is set up with the times at which each boundary is encountered. The perigee point is the first (time=0) and last point (time=period) encountered. The algorithm in SD steps from the lowest latitude boundary to the highest boundary finding the times for each boundary and setting up an index array to specify which altitude bin the object passes through next. The

boundaries and bins are illustrated below for a simple orbit which is only in one altitude bin, and also a more complicated encounter scenario. The figures label the: altitude boundary numbers, the altitude bins numbers, and a boundary encounter number. Notice that the first and last encounter occur at the perigee. Also for the second figure, there is a time around the perigee and apogee that the object is not within the altitude region of interest. This is accounted for in the implementation. See Figure 2.
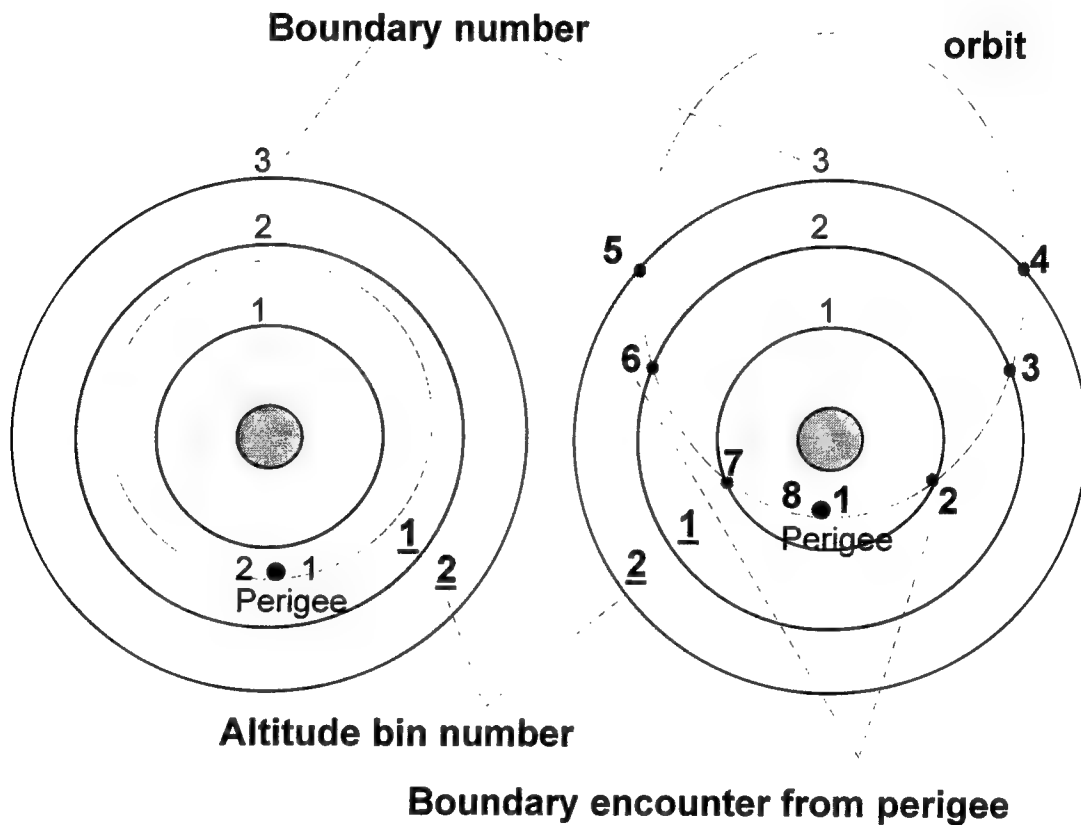

Figure 2 -- Method for determining radial boundary encounters

2) Latitude boundaries - The arrays for the latitude boundary times and indices is analogous to the altitude arrays, but is more complicated due to the lack of symmetry with regard to the time of latitude boundary encounter. The procedure for finding the times and indices of all latitude boundaries and the perigee points is complicated by finding the position along the orbit for a particular boundary. Proceeding from the minimum latitude bin to the maximum latitude bin, one can find the time to a specific boundary by determining the argument of latitude (u) from Eqn. 3. The inverse sine function returns a value between $-\pi/2$ and $\pi/2$ (-90 to 90 degrees).

Hence, this is always an ascending pass of a latitude boundary. The other argument of latitude for the same latitude is found by u2 = π - u1, where u2 is the descending argument of latitude and u1 is the ascending one. A general prograde orbit case is illustrated below in Figure 3.
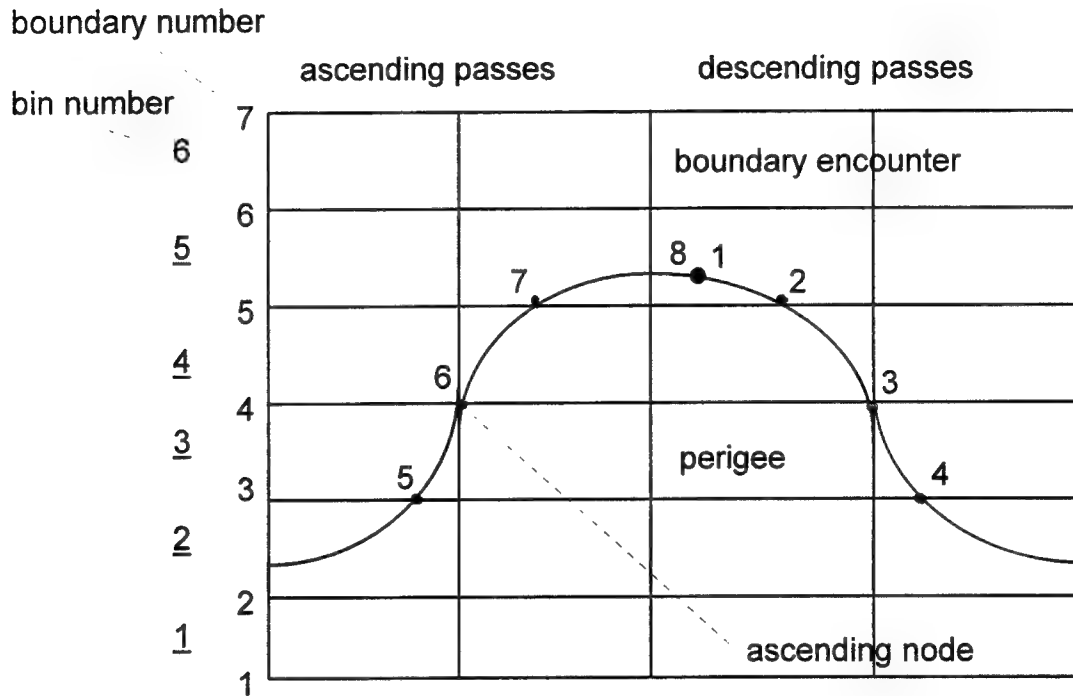
Figure 3 -- Method for determining latitude boundary encounters.

3) Right ascension (longitude) boundaries -- The times and indices for the longitude arrays are found by stepping through the right ascension boundaries from the perigee to 2π (360) and then from 0 to perigee. The argument of latitude can be found for a given right ascension since the longitude of the ascending node is known, and thus the time can be found. For a retrograde orbit the stepping is first from perigee to 0 and then from 2π to perigee. Figure 4 illustrates the boundaries and bins for the longitude.
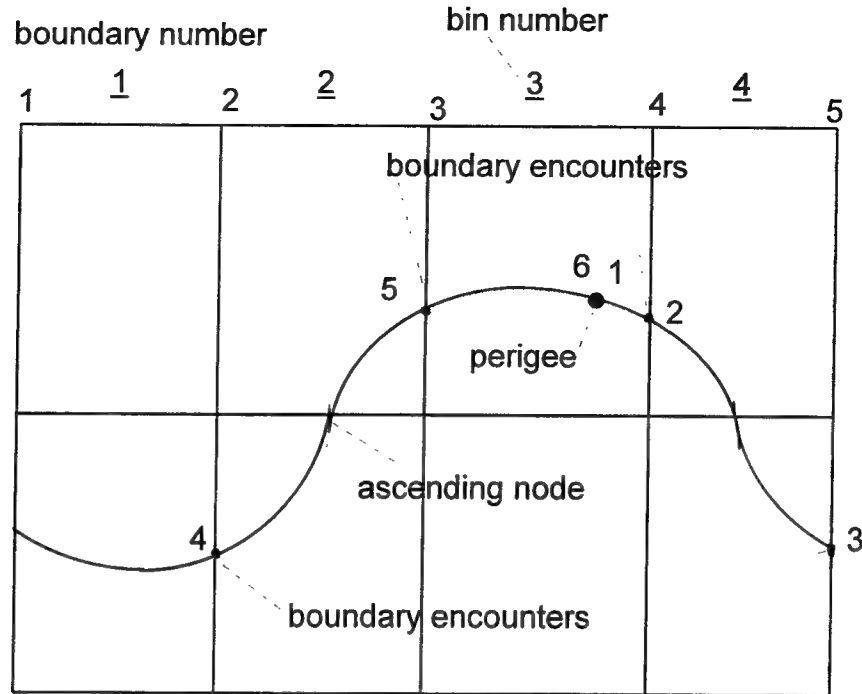
Figure 4 -- Method for numbering longitude boundary encounters.

Spatial Density Calculations With Unknown Orbital Parameters

In the cases for which the ascending node and argument of periapse are not known or changing rapidly. One should use a different method to calculate the average spatial density. This is because one cannot step around the orbit if the periapse is unknown. Instead the spatial density can be calculated using functions for each of the 3 dimensions in a similar method to Kessler's.[10]

$$SD = F(r) + G(\beta) + H(\lambda) \tag{11}$$

In Eqn. (11) F(r) is determined by the percentage of an orbit spent in a radial band using Eqns. (2) and (3). $G(\beta)$ is the percentage of the orbit at a latitude band based on the arc length of the orbit in the band and the mean motion. The last term, $H(\lambda)$ is the fraction of the orbit within a longitude bin and is linear with the bin width.

Results

The method to determine spatial density described in this paper has been incorporated into the DEEP model to calculate the spatial density and collision rates in LEO. The spatial density of the December 1994 tracked object population is presented as a function of latitude and longitude for the altitude ranges 400-500 km and 900-1000 km. This is done to illustrate the significant variance in the spatial density, not only in latitude as expected, but also in terms of the longitude.

Figures 5 and 6 show the instantaneous spatial density of tracked objects between 900 and 1000 km altitude. It can be seen that there are variations within both the longitude and latitude. The latitude variation with higher spatial densities at the higher latitudes is expected because the orbits tend to spend a larger fraction of their orbits at latitudes close to the orbital inclination. This can be visualized by examining a ground track of any space mission. The spatial density variation due to longitude, on the other hand is not really expected due to the precession of the ascending node. In this figure the log of the spatial density has variations reflecting almost an order of magnitude differences between points, with the longitude variation being a factor of 2 to 5. While this may not seem significant, it can affect the hazard to space assets over short time frames.

Figure 7 is a plot of the instantaneous spatial density in the 400 to 500 km altitude range. This range has a significantly lower number of resident objects and hence a lower spatial density. In this snapshot of the orbit, there are some regions of very low spatial density which are cut off on the graph.
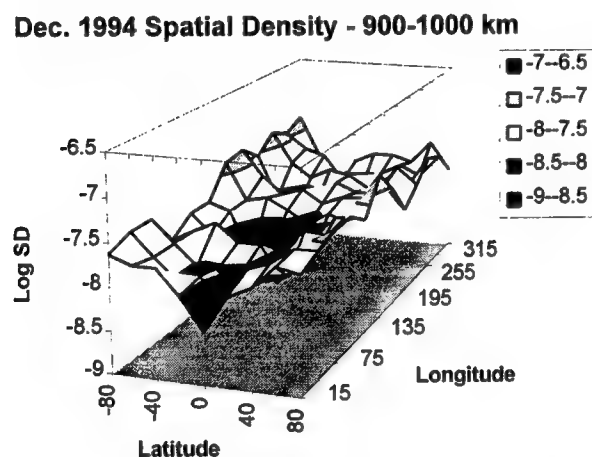
Figure 5 -- The instantaneous spatial density for tracked objects of December 1994.

Dec. 1994 Spatial Density - 900-1000 km
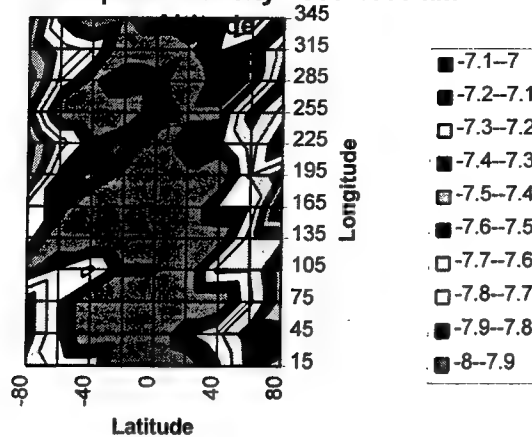
Figure 6 -- Contour plot of the log of the instantaneous spatial density for the tracked December 1994 population. Structure is seen within the longitude dimension.
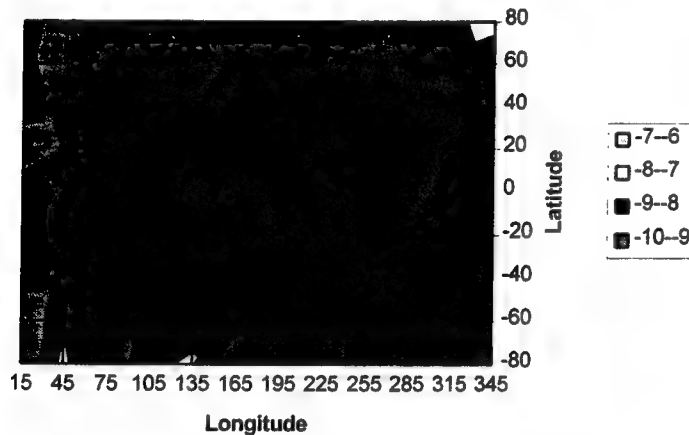
Dec. 1994 Spatial Density - 400-500 km Altitude

Figure 7 -- The log of the instantaneous spatial density of tracked objects on Dec. 15 1994. This contour plot is for the altitude range of 400-500 km.

Conclusions/Summary

The Debris Environment and Effects Program (DEEP) is being developed at the Phillips Laboratory for the purpose of estimating the effects of the debris environment and its evolution. This paper presents recent developments in the spatial density and collision rate algorithms for the DEEP model. The method for determining the spatial density as a function of radius, latitude and longitude have been developed in detail and examples of its implementation shown. The

spatial density routines add a new level of flexibility to space debris hazard assessment. Future developments in determining more accurate values for the impact velocity are continuing.

References

1. Maethner, S.R., A.E. Reinhardt, and L.O. Anderson, "Report on USAF Space Debris Phase One Study," Phillips Laboratory Technical Report PL-TR-94-1042, June 1994.

2. Maethner, S.R., D.B. Spencer, and A.B. Jenkin, "The Air Force Space Debris Research Program," Paper AAS 95-198, AAS/AIAA Spaceflight Mechanics Meeting, Albuquerque, NM, February 13-16, 1995.

3. Spencer, D.B., S.R. Maethner, A.J. Shubert, and K.W. Yates, "The Debris Analysis Workstation: From Concept to Reality," Paper SPIE 2483-10, 1995 Symposium on Optical Engineering and Dual-Use Photonics, Orlando, FL, April 17-21, 1995.

4. Madler, R.A., "Sensitivity of the Near-Earth Orbital Debris Environment to Satellite Fragmentation Parameters," Ph.D. Dissertation, University of Colorado, Boulder, Colorado, 1994.

5. Eichler, P., "Kollisionen zwischen Raumflugobjeckten auf Erdumlaufbahnen und dabei Mogliche Kettenreaktionseffekte," Doctoral Thesis, Technical University of Braunschweig, Verlag Shaker, Aachen, 1993.

6. Eichler, P. and R.C. Reynolds, "Mid- and Long-Term Debris Environment Projections Using the EVOLVE and CHAIN Models," in *Space Environmental, Legal, and Safety Issues*, Proc. SPIE 2483, 1995, pp. 100--113.

7. Yates, K.W., S.R. Maethner, and D.B. Spencer, "Space Debris Environment and Effects Modeling," Paper AIAA 95-0663 33rd Aerospace Sciences Meeting, Reno, NV, January 9-12, 1995.

8. Spencer, D.B., S.R. Maethner, K.W. Yates, and R.A. Madler, "Program DEEP: A Modeling Approach for Estimating the Space Debris Environment and its Effects," to be presented at the 46th Congress of the IAF, Oslo, Norway, October 2--6, 1995.

9. Madler, R.A., and R.D. Culp, "Calculating Collision Rates, Intersection Angles and Relative Velocities Between Intersecting Orbits," Paper AIAA 94-3740, Presented at the AIAA/AAS Astrodynamics Conference, Scottsdale, AZ, August 1-3, 1994.

10. Kessler, D.J., "Derivations of the Collision Probability Between Orbiting Objects: The Lifetimes of Jupiter's Outer Moons," *Icarus*, Vol. 48, 1981, pp. 39--48.

11. Opik, E.J., "Collision Probabilities with Planets and the Distribution of Interplanetary Matter," *Proceedings of the Royal Irish Academy*, Vol. 54, Section A, April 1951, pp. 165--199.

12. Wetherill, G.W., "Collisions in the Asteroid Belt," *Journal of Geophysical Research*, Vol. 72, No. 9, 1967, pp. 2429--2444.

13. Greenberg, R., "Orbital Interactions: A New Geometrical Formalism," *The Astronomical Journal*, Vol. 87, No. 1, January 1982, pp. 184--195.

14. Rossi, A., and P. Farinella, "Collision Rates and Impact Velocities for Bodies in Low Earth Orbit," *ESA Journal*, Vol. 16, 1992, pp. 339--348.

15. Bottke, W.F., Jr. and R. Greenberg, "Asteroidal Collision Probabilities," *Geophysical Research Letters*, Vol. 20, No. 10, May 21, 1993, pp. 879--881.

16. Farinella, P. and D.R. Davis, "Collision Rates and Impact Velocities I the Main Asteroid Belt," *Icarus*, Vol. 92, 1992, pp. 111--123.

17. Bottke, W.F., Jr., M.C. Nolan, R. Greenberg, and R.A. Kolvoord, "Velocity Distributions among Colliding Asteroids," *Icarus*, Vol. 107, 1994, pp. 255--268.

18. Brooks, D.R., G.G. Gibson, and T.D. Bess, "Predicting the Probability that Earth-Orbiting Spacecraft will Collide with Man-made Objects in Space," *Space Safety and Rescue*, Ed. Phillip Bolger, in Science and Technology Series, Vol. 37, AAS Publications, San Diego, 1975, pp. 79--139.

19. McCarter, J.W., "Probability of Satellite Collisions," NASA TM-X-64671, June 8, 1972.

20. Kessler, D.J., R.C. Reynolds, and P.D. Anz-Meador, "Orbital Debris Environment for Spacecraft Designed to Operate in Low Earth Orbit," NASA TM 100471, April 1989.

21. Nauer, David. "SMART Catalog Public/Analyst Data Density Calculations," Teledyne Brown Engineering Technical Report No. CS88-LKD-005, April 1988.

EVALUATION OF RIB/SKIN FRACTURE IN COMPOSITE ISOGRID

Arup K. Maji

Associate Professor

Department of Civil Engineering

University of New Mexico

Albuquerque, NM 87131

.

Final Report for :

Summer Faculty Research Program

Phillips Laboratory, PL/VTSC, NM

EVALUATION OF RIB/SKIN FRACTURE IN COMPOSITE ISOGRID

Arup K. Maji

Associate Professor

Department of Civil Engineering

University of New Mexico

Albuquerque, NM 87131

Abstract

Fracture mechanics was used to model failure of the rib/skin connection of composite 'isogrids'. The energy derivative method was used to determine fracture parameters using a general purpose finite element (FEM) program, with an orthotropic material model. The results of the FEM study were correlated with experimental data obtained from flexure tests. The paper explores the concepts of linear elastic and nonlinear fracture as applicable to these graphite/epoxy composite structures. It provides the designers and users of the isogrid with a methodology for quanitfying structural degradation as a function of the rib/skin disbond.

Electronic Shearography (ES) and Environmental Scanning Electron Microscopy (ESEM) were used to observe crack propagation and fracture mechanisms respectively. The applicability of electronic shearography for local and global nondestructive evaluation (NDE) of isogrid structures was also evaluated.

EVALUATION OF RIB/SKIN FRACTURE IN COMPOSITE ISOGRID

Arup K. Maji

INTRODUCTION

In the early 1970s, McDonnell Douglas pioneered the design of a aluminum structure comprising of a skin, stiffened by a system of grids or lattice (NASA, 1973; Slysh et.al., 1975). A simple triangular system of stiffening elements at 60^0 orientations provided an overall isotropic behavior; hence the name 'isogrid' (Figure 1). This led to the extensive use of aluminum alloy isogrids as the payload shroud and the interstage of the titan and delta spacecrafts. Rhodes and Mikulas (1975) explored the use of composite materials for manufacturing the next generation of isogrid structures. The pioneering work at McDonnell Douglas led to a patent on the manufacture of light-weight graphite/epoxy shell structures (Kam et al. 1978). A number of programs on composite lattice structures ensued in the US and the former Soviet union over the next decade.

Reddy et al. (1985) compared the isogrid with other lattice reinforcement schemes (orthogrid and generalized orthogrid), and concluded that the isogrid was the most attractive due to its damage tolerant characteristics. The redundancy of the rib stiffening system was experimentally proven by cutting individual ribs and exploring buckling loads, redistribution of stresses, etc. (Reddy et al. 1983; Ambur and Rehfield, 1991). Design was governed primarily by the local buckling of the various structural elements that make up the isogrid, such as the rib and skin system, and global buckling of the structural system being designed. Composite isogrids pose additional design problems due to material anisotropy. Mukherjee (1988) discussed the formulation of curved shell elements for analyzing isogrids of complex shapes with the finite element method. The variability in laminae stacking sequence, rib spacing, etc. were subsequently explored with sophisticated finite element analyses (Gurdal and Grall, 1994).

Isogrid fairings will be used in spaceships flown within the next year. Advanced space applications of isogrids are currently being pursued for solar array panels and components of satellites (NASA, 1995).

RESEARCH OBJECTIVES

Most investigators including Kouri et.al (1993) have reported that the separation of the rib/skin connection, followed by the local buckling of individual ribs is the critical failure mechanism in isogrids. Disbonds can develop during manufacture, or due to in-service loads. The technical issues addressed in this paper are :

i)   what are the structural and material attributes leading to fracture ?

ii)   when is a disbond likely to propagate leading to failure of an isogrid structure ?

iii)   can electronic shearography detect and quantify a disbond of significant size ?

FRACTURE MECHANICS

A review of linear elastic fracture mechanics (LEFM) for isotropic materials was provided by Tada et al. (1973), and form the basis for this project. Since few materials if any are perfectly brittle, concepts of nonlinear fracture are under investigation (Lemaitre and Chaboche, 1990).

Energy Release Rate Criteria :

The Energy Release Rate G was used here as the fracture parameter. According to linear elastic fracture mechanics (LEFM), a crack propagates when G = Gc, which is a material property. G is related to the stress intensity factors K_I and K_{II} by the relation :

$$G = \frac{1}{E'}\left(K_I^2 + K_{II}^2\right) , \text{ where E' } = \text{ E (plane stress) or } \frac{E}{(1-\gamma^2)} \text{ (plane strain)} \qquad (1)$$

The use of G rather than K is conceptually simpler for the following reasons :

i)   Failure in composites is very often in mixed mode, and it is not easy to separate the interaction between K_I and K_{II}, or to have available a mixed-mode failure criteria. The manufacturer of the epoxy provided a value for G, and no other fracture related property.

ii)   For an interface crack, the theoretical solution for near-tip stresses σ, are oscillatory, and the normal and shear stresses are coupled (Rice, 1988) :

$$\left(\sigma_{yy} + i\sigma_{xy}\right)_{\theta=0} = Kr^{i\varepsilon} / \sqrt{(2\pi r)} \qquad (2)$$

where ε is a bimaterial constant. The physical significance of such near-tip stress field is not well established. These considerations led to our use of G, rather than the stress intensity factors as the relevant fracture parameter.

FINITE ELEMENT ANALYSES

The finite element method (FEM) using an energy derivative approach (Parks, 1974) was used to determine G. According to this method, G is equal to the change in the total strain energy corresponding to an incremental extension of a crack, while the load (P) is held constant :

$$G(a) = \frac{\partial \Pi}{\partial a} = -\frac{1}{2} \frac{\partial \{[u][K][u]\}}{\partial a} \qquad (3)$$

where a is the crack length, Π is the total strain energy of the specimen obtained from the displacement [u] and stiffness [K] matrices. Therefore the change in the total strain energy of the specimen is to be determined at two slightly different crack lengths.

A commercially available finite element software, 'IDEAS' (SDRC, 1993) was used because of its ability to incorporate orthotropic material model and its postprocessing capabilities. Since this software did not have any specific fracture mechanics capability (such as singular elements), G was evaluated from the energy derivative approach discussed earlier (equation 3).

Study of Accuracy :

A Single-Edge-Notched geometry with isotropic material property was first analyzed with this approach (Figure 2). The analytical solution for K, and hence G (see equation 1), was available for this from Tada et al. (1973) geometry (to an accuracy of better than 0.5% for any a/b):

$$K_I = \sigma \sqrt{\pi a} F(a/b)$$

$$F(a/b) = \sqrt{\frac{\tan \psi}{\psi}} \frac{0.752 + 2.02(a/b) + 0.37(1 - \sin \psi)^3}{\cos \psi} \qquad (4)$$

where $\psi = \pi a/2b$.

After a convergence study was performed, a 1800 element mesh was eventually used to model a 3"x6" Single Edge Notch (SEN) specimen with a 0.9" crack (Figure 2). Strain energy (Π) output of the software was erroneous and had to be determined separately from the external work done, from the relation :

$$\Pi = \frac{1}{2} P \delta \tag{5}$$

Unit normal distributed load was applied along the outer edge. Accuracy of the displacement output were therefore verified independently by analyzing the same mesh with the MSC-NASTRAN code. The values of G were obtained from FEM analyses output of the total strain energy, Π. Figure 3 compares the FEM G values to the theoretical solution ($= 2.58 \times 10^{-7}$ lb/in) for various small crack increments δa in the FEM analysis, for a 200 elemnt mesh and a the 1800 element mesh.

T-Beam Model :

The T-beams tested in our laboratory (discussed later) were modeled next using a 2-dimensional mesh with a different thickness for the rib and the skin sections (Figure 4). Orthotropic material properties were used, different for the rib (web) and the skin (flange), as discussed in the next section. A convergence study was performed by progressively refining the mesh from 48 to 2048 elements. Four noded rectangular 'thin shell' elements were used to generate the initial mesh. Table 1 lists the y-coordinates of the bottom of each column of elements. All elements had the same dimension in the x direction (128 elements over 2", ie., 7.8125×10^{-3} in. wide). This mesh was analyzed under three-point-bending load and the results were verified by checking the deflections and stresses through hand calculations. The variation in rib thickness (discussed later) was simulated by changing the thickness of the elements simulating the ribs from 0.05" to 0.03" progressively (Table 1). Crack surfaces were then introduced by separating the nodes on the two sides at the appropriate locations. Finally, the T-beam geometry with cracks of 1" and 1/2" length (Figure 4) at the rib-skin interface were analyzed to determine the load-point displacements (Figure 5). From this data, G was evaluated using equations 5 and 2.

Orthotropic Material Properties :

The transversely isotropic materials such as the graphite/epoxy composites used here, have 5 independent constants (Daniel and Ishai, 1994) :

Properties of the rib in the fiber direction (1) is different from those in the other two directions (2 and 3) (Figure 1). However, for the [0/+60/-60]s lay-up of the skin, the skin is symmetric and isotropic in its plane while the out-of-plane direction has different properties, dominated by the matrix. The following material properties were available from the manufacturer (ICI Fiberite, Tempe, AZ) : E_{11} = 24.4 Msi., E_{22} = E_{33}= 1.1 Msi., G_{13} = G_{12} = 0.80 Msi., v_{13}= 0.30. Assuming that the fibers are perfectly bonded (continuity of strain), it can be shown that v_{23} = 0.30. These properties were applicable to the unidirectional rib of the isogrid. For the [0/+60/-60]s skin lay-up, calculations resulted in the following 5 independent properties for the skin : E_{11} = E_{22}= 8.9 Msi., E_{33}= 1.1 Msi., G_{13} = G_{23} = 0.55 Msi., v_{13} = 0.30, v_{12} = 0.30. Note that in the isotropic 1-2 plane, G_{12} = $E_{11}/2(1+ v_{12})$ = 3.42 Msi., is not an independent constant. The derivation of G_{13} = G_{23} and v_{13} = v_{23} were done as per the text by Daniel and Ishai, (1994).

EXPERIMENTAL PROGRAM

The specimens were cut out of a 18" square isogrid panel made of IM7 graphite fiber and 977-2 epoxy prepregs and tows. The isogrid manufacture process is described by Koury et al. (1993). Diamond cutting blade was used to cut rib-skin specimens (Figure 1), 2" in span. Rib and skin thicknesses were 0.05" and 0.02" respectively. Rib height was 0.5" and the skin width was 1.0".

Three point bending tests were conducted with a screw type loading device without any close-loop control. The loading fixture was mounted on an optical table to facilitate examination with electronic shearography. Displacements were measured with a dial gage (0.001" sensitivity). Loads were measured with a 200 lb. capacity load cell and associated instrumentation. The specimens were cut out of an isogrid panel fabricated at the Phillips laboratory, with a diamond cutting blade. Artificial cracks, 1" and 1/2" long were introduced at the rib-skin connection by a rotary cutting blade with a diamond

abrasive wheel. The specimens were supported at the ends by embedding the ends into a Plexiglas cube with slits that matched the rib, and using epoxy to hold the rib to the Plexiglas. This was necessary to prevent local rib-buckling at the supports. Loading was applied through a cylindrical piece of steel, 1" in diameter. Three samples of each of the three types (no crack, 1/2" and 1" crack) of specimens were tested.

Figures 6a and b show representative data from the three-point-bending tests on the 1/2" and 1" crack specimens respectively. All the specimens tested failed at the rib-skin connection, and from the tip of the artificial notches as expected. Some secondary shear cracking (parallel to the span) was observed in the ribs, especially for the specimens without artificial cracks. Figure 7 summarizes the failure loads from all the tests.

G per unit load for the 1/2" and 1" crack specimens (Figure 5) were 0.0199 and 0.065 /in respectively. The manufacturer provided a Gc value of 2.1 lb.in/in^2 for the resin. Based on the LEFM energy release rate criteria (crack propagates when G = Gc), the failure loads for the 1/2" and 1" specimens were calculated to be 106 and 32 lbs respectively. This agrees reasonable well with the test data in Figure 7, although LEFM predicts a sharper decrease in strength than was observed experimentally. The reasons for this behavior is due to nonlinear aspects of fracture such as plasticity (in metals) and mirofracture (in composites and ceramics) and has been extensively discussed in the literature (Lemaitre and Chaboche, 1990).

Rib Thinning :

Upon measuring the ribs upon fracture with vernier callipers, it was observed that the rib thickness varies consistently from 0.05" at the outer edge to 0.03" at the skin. This decreased thickness at the skin significantly decreases the shear strength in this most critical location. It is presumed that the decreased thickness is a result of the thermal expansion of the rubber molds used during isogrid manufacture. Our suggestions are therefore to either taper the slits in the rubber mold (expensive tooling) to accommodate the expansion or use a material with lower CTE (coefficient of thermal expansion). Such mandrels, using polyurethane foam are currently under investigation in a joint project involving the Phillips laboratory and Lockheed Martin Corp. (NASA, 1995).

ESEM Study :

An environmental scanning electron microscope (ESEM) from Electroscan corp. was used to inspect the composite materials before and after loading, to examine the microstructural defects and to identify the failure mechanisms. The ESEM does not require the vacuum necessary for conventional electron microscopy. Hence, it is not necessary to apply any charge dissipative coatings on the. The specimens were polished on a diamond abrasive wheel to 1 micron prior to the ESEM study.

The matrix appears to coat and bond well with the fibers (Figure 8a). Brittle fracture of the fibers is evident from the flat fracture surfaces (Figure 8b). There were no visible voids, possibly due to resin movement during curing. It was possible to see the variations of the fiber volume fraction through the cross-section of the rib (Figure 8c). Some areas devoid of any fibers is also visible, providing for weak zones susceptible to shear cracking (Figure 8d). These weak shear planes sometime cause rib failure away from the rib/skin interface, explaining the scatter in some of our test data.

Electronic Shearography (ES)

Shearography is a laser based interferometry technique developed in the early 80s using conventional films and later coupled with electronic (digital) data acquisition and image processing (Hung, 1989). The technique is full-field, noninvasive, and relatively robust compared with other interferometry techniques. The ES system used here has a birefringent crystal which separates one beam coming from the object into two, orthogonally polarized beams (Figure 9). The image acquired by the ccd (charge coupled device) camera is therefore a superposition of two images coming from the object that are separated by a distance δx by the birefringent crystal. Since a coherent laser beam is used (Helium-Neon, with $\lambda = 0.63\mu$), the images contain numerous black dots called 'speckles'. To perform electronic shearography, a reference image is first stored in the image analysis system. All subsequent images are automatically subtracted from this reference image. This digital subtraction generates an interference fringe pattern of the type reported here. Therefore the ES system provides displacement gradient, which could

be strain or the slope of the deflection curve, depending on the setup used. This technology has been applied to detect disbonds in various composite panel structures (Tyson, 1991).

The setup used involved a 35mw He-Ne laser, with associated optics. A Shearography camera (model SC4000) and control system (model CCU4000) from Laser Technology Inc. were used. Digital image processing was performed on a personal computer containing a frame grabber and the 'Imaster' image processing software.

Typical shearography images are shown in Figures 10a and b. These depict the strain concentrations at the tips of the 1/2" and 1" cracks respectively. The location of these strain concentrations do not change during loading up to failure, implying that the cracks do not propagate until the maximum load, when they propagate in an unstable manner.

Detecting Disbonds with Shearography :

A second isogrid panel 18" square was inspected with the ES system to study the feasibility of detecting rib disbonds. The area inspected was reduced to 6" square at one time. Otherwise the speckle size was too small due to limitations of our hardware. The panel was very reflective and the variation in amplitude due to the speckles was smeared by the light directly reflected.

The isogrid panel was clamped along the top and bottom edges and an out-of-plane point force was applied at different locations. Figure 11a shows the circular fringe pattern when the loading is at the center of the pockets. When the loading is close to a well-bonded rib, no fringe pattern is observed at the same load level. When the loading is increased significantly (factor of 5) a fringe pattern can be observed (Figure 11b), that terminates at the rib location. The bottom edge of these fringes are straight due to the rib transferring the load and therefore preventing the deformation gradient from showing up in the adjacent pocket (triangle). While there were no broken ribs in the specimen it is expected that a separated rib will result in a fringe pattern similar to that in Figure 11a rather than that observed in Figure 11b. This would allow an inspector to evaluate each rib in an isogrid system with electronic shearography.

Consequence of Rib/Skin Disbond :

Disbond reduces the buckling load of both the rib and the skin. The buckling of these constituents were summarized by Koury et al. (1993). If the rib is considered simply supported at the three edges where it is connected, its buckling load is given by :

$$N_x = D_{11} \frac{\pi^2}{a^2} + 12 \frac{D_{33}}{d^2}, where\ D_{11} \equiv \frac{E_x b^3}{12(1 - \gamma_{xy} \gamma_{yx})}, D_{33} = \frac{G_{xy} b^3}{12} \tag{6}$$

where d is the rib height (=0.5") and b is rib width (0.05"). When the rib separates from the skin, the buckling load reduces to only the first term of equation 6. For the isogrid studied here ($a = 3.6$"), the second term is twice as large as the first. Therefore rib/skin debonding will reduce the rib's buckling load to 1/3 of its original value.

Skin buckling load is determined by the largest circle that can be inscribed within the simply supported boundaries. Therefore for the undamaged case, the largest circle inside the triangular ribs has a diameter of $a\tan30^0$. When a rib is separated, the largest circle that can be inscribed inside two adjacent triangles has a diameter of $a\cos30^0$. The buckling load varies inversely with the square of the circle diameter. Therefore it will decrease the skin buckling load to 44% of the original value.

CONCLUSIONS

LEFM was resonably good at predicting the propagation of rib/skin disbonds. This was experimentally verified for a rather complex state of stress concentration. Aspects of nonlinear fracture that help understand the experimental data were discussed. The energy derivative method was used to determine the strain energy release rate for a composite T-beam. The accuracy of various FEM meshes with orthotropic material properties was studied. The consequence of rib/skin disbond in the isogrid system was estimated, based on the primary buckling modes of the individual components.

Various causes of fracture and material variability based on microstructural characteristics were studied with an environmental scanning electron microscope. The ability of electronic shearography to inspect fracture propagation in real time was explored. The ability of the ES method to deveop into a NDT tool for inspecting rib/skin disbonds was demonstrated.

REFERENCES

1.	Ambur D.R. and Rehfield L.W., Effect of Stiffness Characteristics on the Response of Composite Grid-Stiffened Structures", AIAA paper no. 91-1087-CP, 1991.

2.	Daniel I.M. and Ishai O., "Engineering Mechanics of Composite Materials", Oxford University Press, 1994.

3.	Gurdal Z. and Grall B., "Buckling Analysis of Geodesically Stiffened Composite Panels with Discrete Stiffeners", J. of Aircraft, V31, No. 5, Sept.-Oct. 1994, pp. 1197-1204.

4.	Hung Y.Y., "Shearography: A Novel and Practical Approach for Nondestructive Inspection", J. of Nondestructive Evaluation", V8, No.2, 1989, pp. 55-67.

5.	SDRC, "I-DEAS", Structural Dynamics Research Corporation, 2000 Eastman Dr., Milford, OH 45150, 1993.

6.	Kam C.Y., Freeman V.L., Penon A.P., "Stiffened Composite Structural Member and Method of Fabrication", US Patent No. 4,086,378, 1978.

7.	Koury J.L., Kim T.D., Tracy J.J. and Harvey J.A., "Continuous Fiber Composite Isogrid Structures for Space Applications", ASM Conference, Long Beach, CA 1993.

8.	Lemaitre J. and Chaboche J.L., 'Crack Mechanics', in "Mechanics of Solid Materials", Cambridge university Press, 1991, ISBN no. 0-521-32853-5, pp. 451-549.

9.	Mukherjee A.K., "Isogrid Shell in Aerospace Structures", IE(I) Journal, V69, Sept. 1988- March 1989, pp. 6-9.

10.	NASA report no. CR124075, prepared by McDonnell Douglas Astronautics Co., 1973.

11.	NASA Technical Memorandum TM X 72771, 1995.

12.	Newman J., "Shearographic Inspection of Aircraft Structure", Materials Evaluation, V49, No.9, 1991, pp. 1106-1109.

13.	Parks D.M., "A Stiffness Derivative Finite Element Technique for Determination of Crack-tip Stress Intensity Factors", Int. J. of Fracture, V10, pp. 487-502, 1974.

14.   Reddy A.D., Rao V.R. and Rehfield L.W., "Continuous Filament Wound Composite Concepts for Aircraft Fuselage Structures", J. Aircraft, V22, No. 3, March 1985, pp. 249-255.

15.   Reddy A.D., Rehfield L.W., Haag R.S. and Widman C.G., "Compressive Buckling Behavior of Graphite/Epoxy Isogrid Wide Columns with Progressive Damage", ASTM STP 808, R. Chait and R. Papirno Ed., 1983, pp. 187-199.

16.   Rehfield L.W., "Studies of Advanced and Composite Structures", AFOSR Report no. 82-1016, 1982.

17.   Rice J. R., "Elastic Fracture Mechanics Concepts for Interfacial Cracks", Transactions of ASME, V55, March, 1988, pp, 98-103.

18.   Rhodes M. and Mikulas M.M., "Composite Lattice Structures", NASA Technical Memorandum, TM X-72771, 1975.

19.   Slysh P, Dyer J.E., Furman, J.H., and Key J.E., "Isogrid Structures", AIAA paper no., 75-816, 1975.

20.   Tada H., Paris P.C. and Irwin G.R., "The Stress Analysis of Cracks Handbook", Del Research Corp., Hellertown, PA 1973.
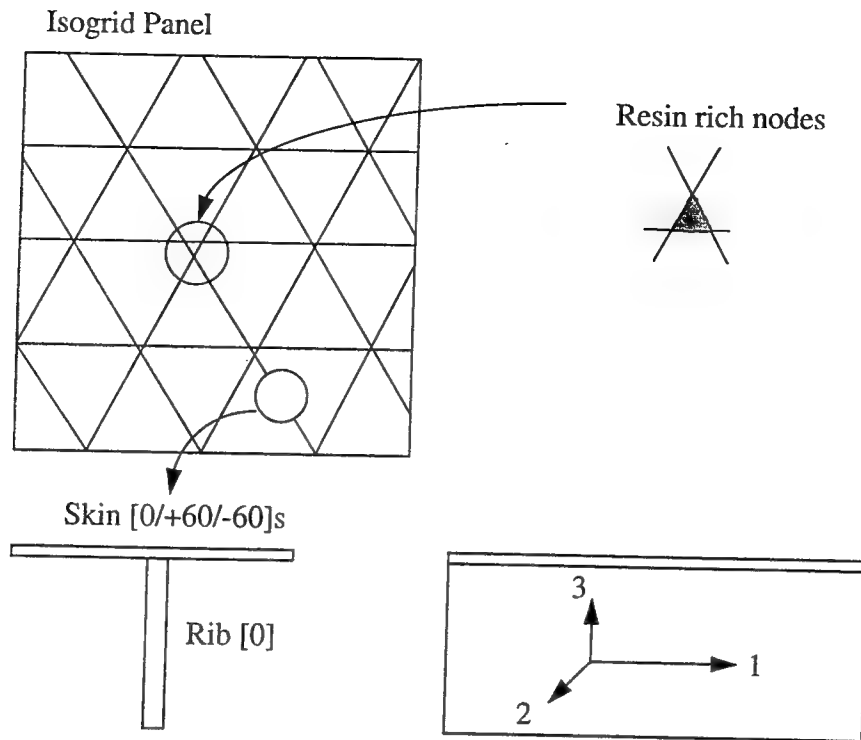
Figure 1. Isogrid Panel, Components and Directions
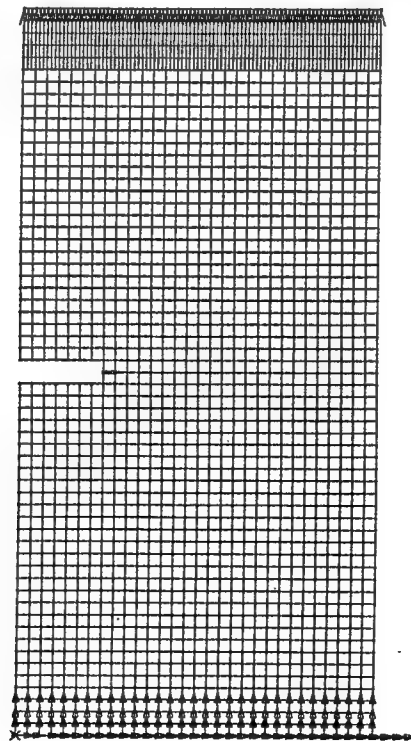
Figure 2. 1800 Element Mesh for SEN Specimen
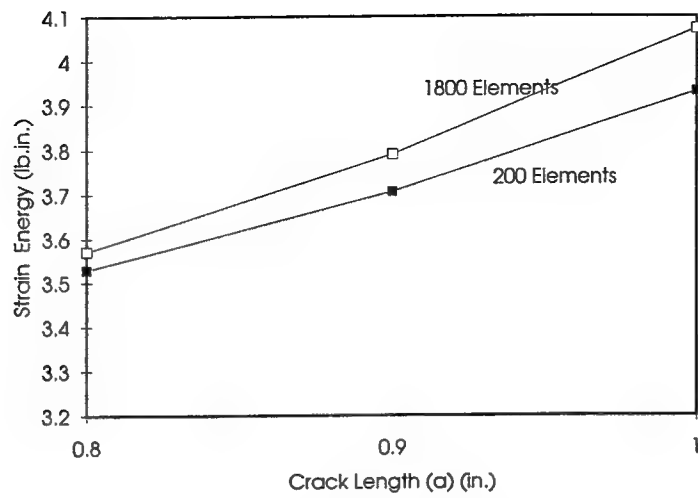
Figure 3. Strain Energy vs. Crack Length 'a'
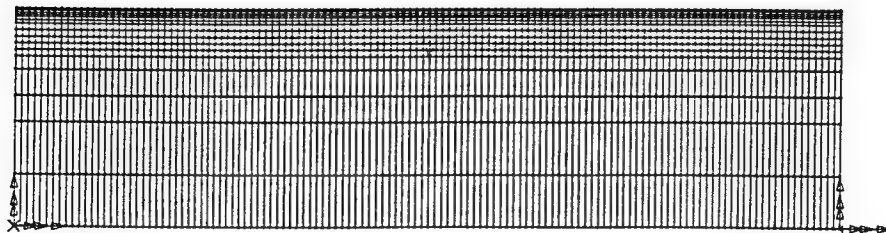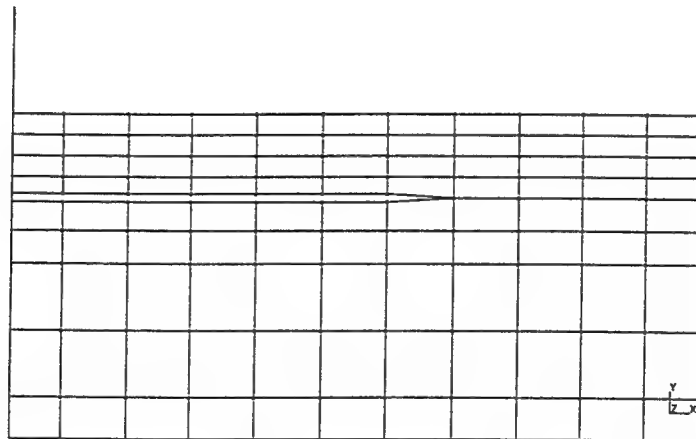
Figure 4. Refined Mesh
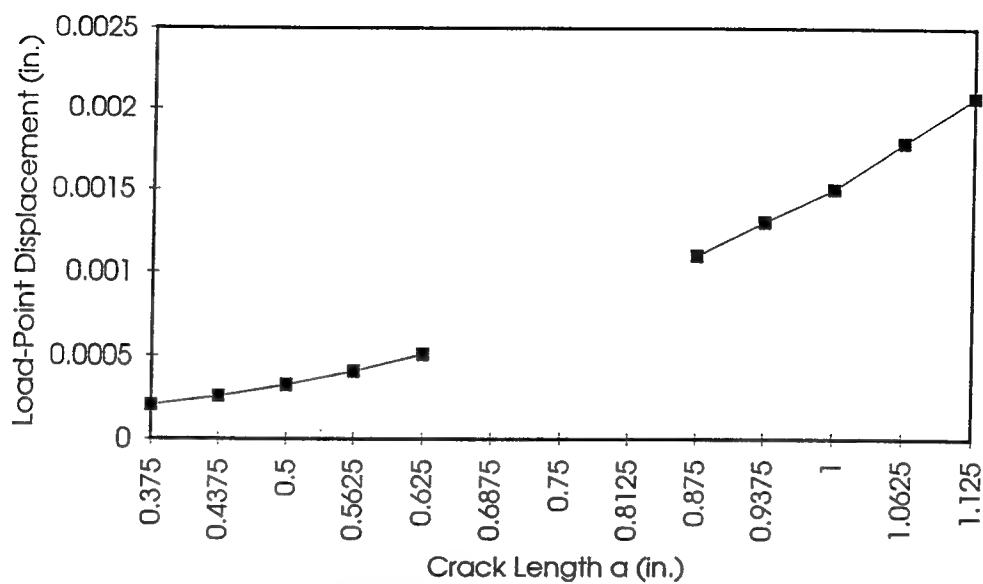
Crack Extension vs. Load-point Displacement				
a	d			
0.375	0.000205			
0.4375	0.000259			
0.5	0.000326			
0.5625	0.000408			
0.625	0.000507			
0.6875				
0.75				
0.8125				
0.875	0.001097			
0.9375	0.0013			
1	0.0015			
1.0625	0.001787			
1.125	0.00207			

Figure 5. Load-Point Displacements from T-beam Analyses
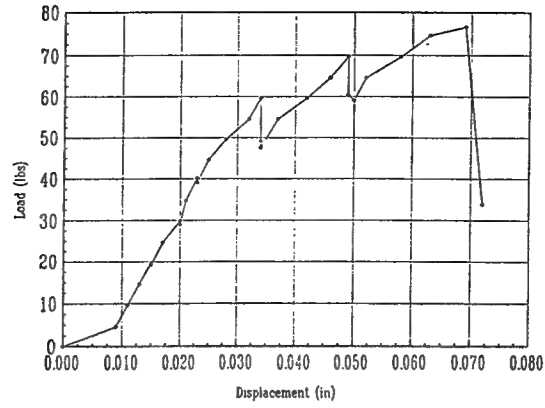
Figure 6a. Load-Displacement for 1/2" Crack Beam
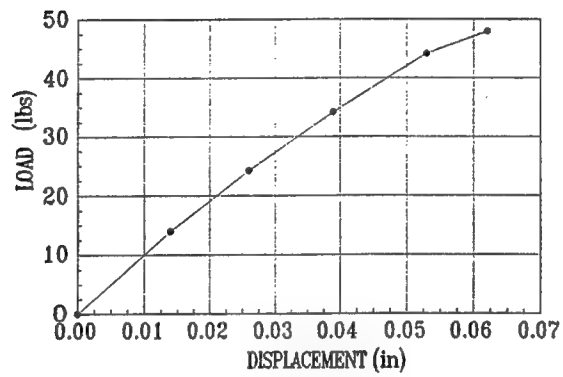
Figure 6b. Load-Displacement for 1" Crack Beam
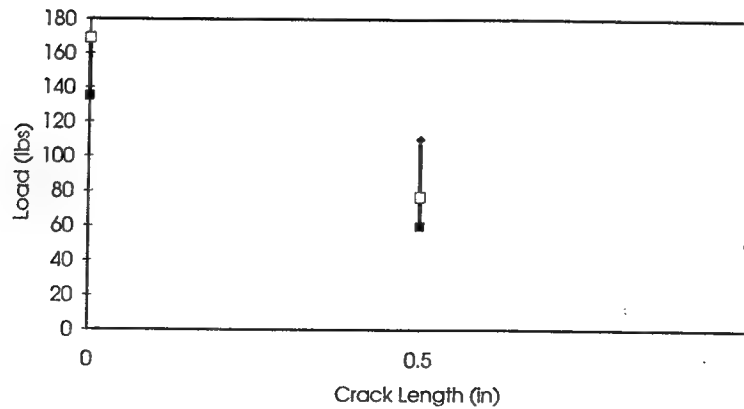
Figure 7. Failure Loads for the Three Types of Beams Tested

25-17

Figure 8a. Epoxy Coating of Fibers   Figure 8b. Brittle Fracture of Fibers
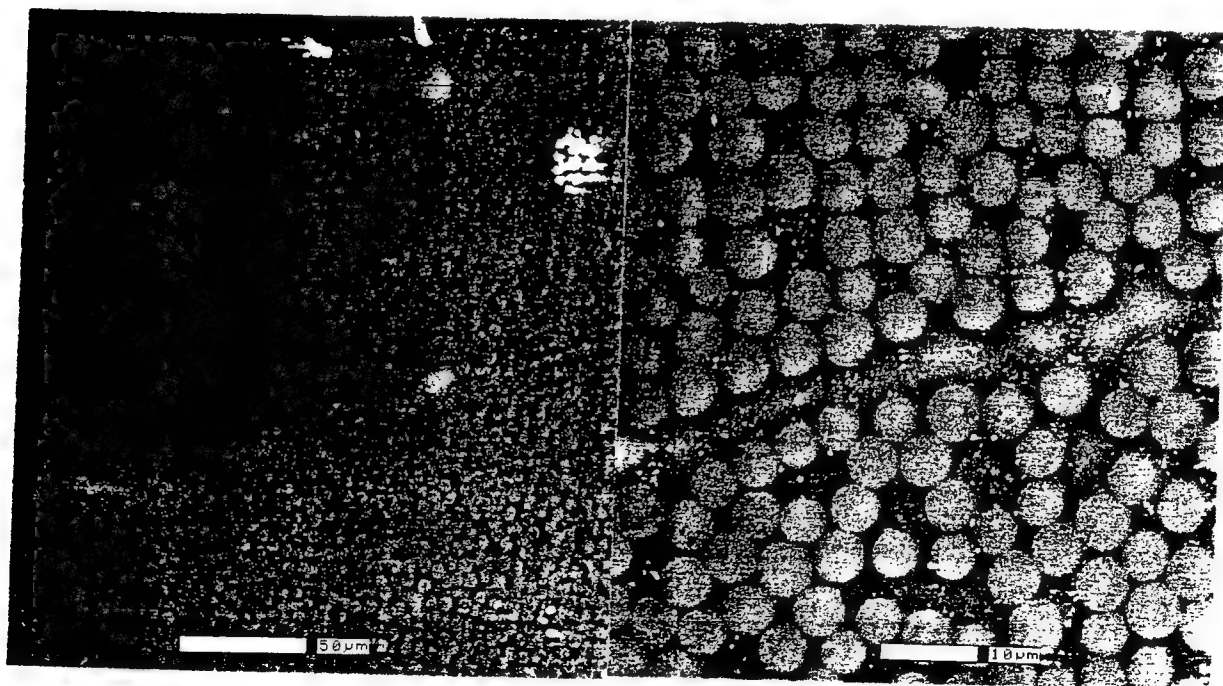
Figure 8c. Variation in the Fiber Volume Fraction   Figure 8d. Weak Plane without Fibers
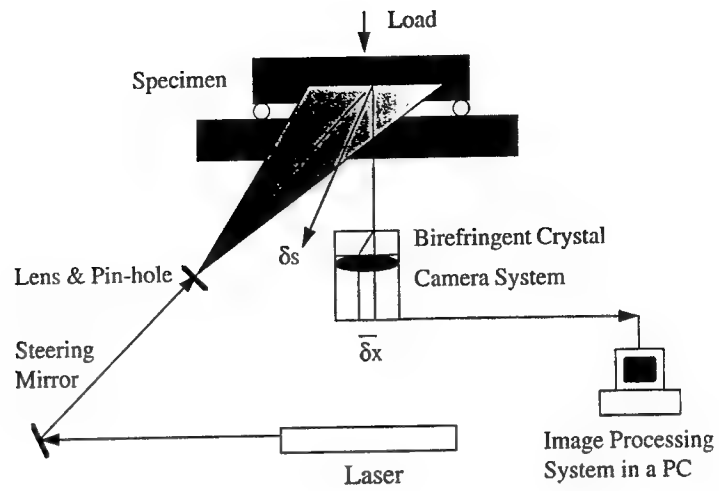
Figure 9. Setup for Electronic Shearography

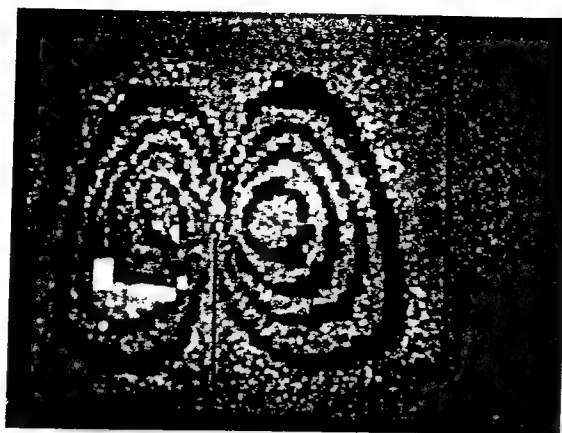Figure 10a. Shearography Image from 1/2" Crack Specimen

Figure 10b. Shearography Image from 1" Crack Specimen

Figure 11a. Shearography Image from Loading at Center of Panel
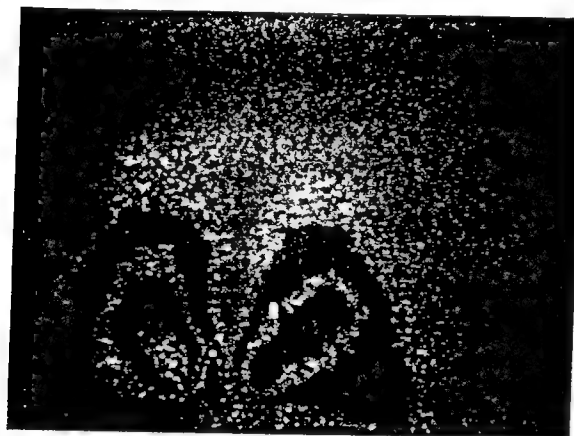
Figure 11b. Shearography Image from Loading at the Rib

Table **1.**

Y Coordinates of the bottom of elements in one column				
Element	Y- value	height (in.)	Thickness (in.)	
1	0	0.125	0.05	
2	0.125	0.125	0.05	
3	0.25	0.0625	0.04	
4	0.3125	0.0625	0.04	
5	0.375	0.03125	0.03	
6	0.40625	0.015625	0.03	
7	0.42187	0.015625	0.03	
8	0.4375	0.015625	0.03	
9	0.45312	0.015625	0.03	
10	0.46875	0.015625	0.03	
11	0.48437	0.007813	0.03	
12	0.49219	0.007813	0.03	
13	0.5	0.05	1	
14	0.505	0.05	1	
15	0.51	0.05	1	
16	0.515	0.05	1	
	0.52			

SYNTHESIS, PROCESSING, AND CHARACTERIZATION OF THERMALLY CROSSLINKABLE THERMOTROPIC COPOLYESTERS

David C. Martin

Assistant Professor of
Materials Science and Engineering and
Macromolecular Science and Engineering

2022 H. H. Dow Building
Ann Arbor, MI 4 8109-2136

with:

Gary Spilman and Tao Jiang
The University of Michigan, Ann Arbor, MI

Patrick T. Mather and Kevin Chaffee
Philips Laboratory, Edwards AFB, CA

Final Report for:

Summer Faculty Research Program
Philips Laboratory

September 22, 1995

SYNTHESIS, PROCESSING, AND CHARACTERIZATION OF THERMALLY CROSSLINKABLE THERMOTROPIC COPOLYESTERS

David C. Martin

Assistant Professor of
Materials Science and Engineering and
Macromolecular Science and Engineering

2022 H. H. Dow Building
Ann Arbor, MI 4 8109-2136

with:

Gary Spilman and Tao Jiang
The University of Michigan, Ann Arbor, MI

Patrick T. Mather and Kevin Chaffee
Philips Laboratory, Edwards AFB, CA

Abstract

The synthesis, processing, and characterization of thermotropic copolymers composed of hydroxybenzoic acid (HBA), hydroxybenzoic acid (HBA), and systematically varying amounts of hydroquinone (HQ) and crosslinkable terephthalic acid (XTA) is described. The XTA monomer contains a benzocyclobutene (BCB) group that lies dormant during synthesis and processing but can be thermally activated to introduce covalent crosslinking between laterally adjacent macromolecules. The XTA-containing HBA/HNA copolymers all remain thermotropically liquid crystalline, and could be processed into oriented fibers by melt spinning. Rheological characterization revealed an increase in the viscosity and a transition from liquid-like to solid-like behavior as the crosslinking reaction proceeds. X-ray diffraction revealed that the microstructure of the XTA containing copolymers was similar to that of neat HBA/HNA, with the extent of crystalline order decreasing slightly with increasing XTA content. TGA experiments show that the HBA/HNA copolymers containing XTA have a decreased onset of thermal degradation temperature, and an increased char content above 700 C.

SYNTHESIS, PROCESSING, AND CHARACTERIZATION OF THERMALLY CROSSLINKABLE THERMOTROPIC COPOLYESTERS

David C. Martin

Assistant Professor of
Materials Science and Engineering and
Macromolecular Science and Engineering

2022 H. H. Dow Building
Ann Arbor, MI 4 8109-2136

with:

Gary Spilman and Tao Jiang
The University of Michigan, Ann Arbor, MI

Patrick T. Mather and Kevin Chaffee
Philips Laboratory, Edwards AFB, CA

Introduction

Thermotropic liquid crystalline polymers (TLCP's) are of scientific and technological interest because of their low melt viscosity, low shrinkage during processing, and excellent mechanical properties. TLCP's have been considered for replacing metals such as aluminum in situations where weight savings are critical to performance. However, TLCP's suffer from a number of technical limitations including poor lateral properties, low compressive strength, high creep, low adhesion, and moderate flame resistance. Also, the highly shear thinning nature of TLCP's during melt processing leads to heterogeneous microstructures including a thin, oriented skin, with considerably less orientation in the middle of the section. In order to achieve the ultimate performance of parts prepared from TLCP materials, it would be advantageous to better control the anisotropy of orientation and mechanical properties.

It has been suggested that chemical crosslinking might be one means by which these current limitations in TLCP performance could be improved. Recently, a benzocyclobutene functionalized variant of terephthalic acid (XTA) has been developed (Walker, Markoski, and Moore, 1992). The advantages of the XTA monomer are that the BCB group is compact, so that upon substitution of TA the mesophase is retained. Furthermore, the BCB group does not react until fairly high temperature (~375 C), so that the polymer can be processed at lower temperatures and then the reactive groups activated at a later stage in the process. Finally, the BCB group crosslinks without liberating small molecules that might lead to voids or porosity. The known disadvantages of BCBs include a reduction in thermo-oxidative stability, and limited availability of the monomer. Also, the influence of BCB modifications on the macroscopic properties and processing of polymers is not yet well established.

In previous work with lyotropic LCP's, we found that the incorporation of BCB groups could lead to improvements in strain to failure by kinking and reduction in creep (Jiang, et al., 1995). However, it was found that kinking persisted even highly

crosslinked PPTA fibers (Jones and Martin, 1995). It was proposed that one means by which a more uniform crosslinking across the section thickness might be achieved is by examining liquid crystalline polymers which could be processed from the melt.

In this paper, we discuss the synthesis and characterization of thermally crosslinkable thermotropic copolyesters. The copolymers we investigated were based on hydroxybenzoic acid (HBA) and hydroxynaphthoic acid (HBA). We chose to examine the 75 HBA/25 HNA composition which corresponds to the commericially important Vectra® copolymers. To the 75 HBA/25 HNA copolymer, we incorporated systematic amounts of the crosslinkable terephthalic acid monomer (XTA) and stroichiometrically balanced equivalents of hydroquinone (HQ). The goal of this work was to establish whether there was an accessible temperature window within which the LCP's could be processed, and if so, what composition of XTA was the most desirable.
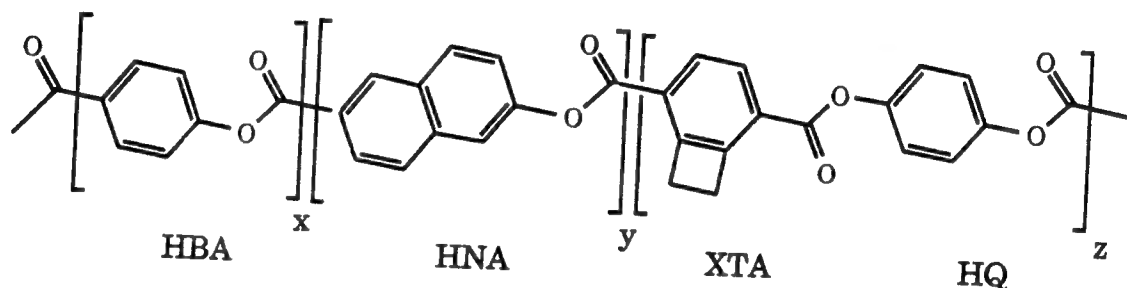
Background

Most thermotropic liquid crystalline polymers studied to date have been aromatic polyesters. Much of the experimental effort has involved copolymers of hydroxybenzoic acid (HBA) and hydroxynapthoic acid (HNA). Copolymers comprised of HBA and HNA form the basis for the Vectra® series of materials which have been commercialized by Hoechst-Celanese, Inc., in Summit, NJ.

The synthesis, processing, and properties of thermotropic liquid crystalline polymers are discussed in considerable detail elsewhere (Jackson and Shaw, 1991; Donald and Windle, 1993). Here, we review the salient features of previous studies as they relate to our research. The synthesis, thermal stability, microstructure (Blackwell, 1983; Cheng, et al., 1989, Windle, Sawyer, et al., 1993; Balta-Calleja, 1993), rheology (Wissbrun, 1981; Kiss, 1986; Larson and Doi, 1991), transesterification (Economy, Blackwell, 1994), molecular weight, foaming (Kalika, 1990; Economy et al., 1994), and low temperature mechanical relaxations (Weinkauf and Paul, 1992) have been studied. The spinning of TLCP's into fiber has been studied by Lee and Dibenedetto (1992, 1993).

Other crosslinkable LCP's have also been studied. Pakula and Zentel examined the rheology of crosslinked liquid crystalline polymers (1991). They found that a minimum in the viscosity could be obtained because of the enhanced mobility of the nematic phase. However, in this system the chemistry of the crosslinking was permanent, and did not change during heating as in our materials.

McGrath et al. have recently described polyimides which could be crosslinked with phenyl-acetylene terminal groups (1995). While not liquid crystalline, these polymers are similar to XTA functionalized materials in that they chemically react at temperatures on the order of 400 C. McGrath et al. found that the complex viscosity of these polymers first decreases and then increases during heating. They observed a relatively low char content at high temperature, but the char content did increase in the crosslinked systems.

Marks has reported on the use of terminal BCB's to crosslink polycarbonates (1994). The introduction of the BCB units caused an increase in the polycarbonate solvent resistance, tensile modulus, yield strength, surface hardness, and a lowering in the elongation to break and 5% weight loss temperature. Marks also found an initial decrease, followed by an increase in the melt viscosity during thermal curing.

A thermally crosslinkable, benzocyclobutene functionalized variant of terephthalic acid (XTA) has been reported by Walker et al. (1992). The ability for XTA and its derivatives to be incorporated into a number of high performance polymers has now been confirmed. The microstructure (Jones et al., 1995) processing, and properties (Jiang et al., 1995) of XTA modified copolymers of PPTA (lyotropic LCP's) has been reported.

While the versatility and utility of BCB functionalization for crosslinking polymers has been confirmed, there remain several outstanding questions which need to be addressed before these materials will be used in engineered structures. Of particular concern is the fact that there have been observations of free-radical formation as a function of time after crosslinking (Mielewski, Bauer, and Martin, 1994).

There is also an interest in evaluating the utility of BCB functionalization for improving flame resistance. The hypothesis is that crosslinking can restrict flow, leading to improved char formation during exposure to elevated temperature. Marks has shown that the addition of BCB groups significantly improves the ignition resistance of polycarbonates (1994).

While it has been demonstrated that BCB moieties can be efficient crosslinking agents, they have also been shown to reduce thermo-oxidative stability. This is further complicated by the fact that the degradation may continue at room temperature over extended periods of time. It is not yet known whether BCB crosslinking can improve mechanical properties to an extent sufficient to overcome these limitations. However, the ability to examine the influence of chemical crosslinking on the macroscopic properties of liquid crystalline polymers makes the ability to control BCB concentration a useful means for obtaining fundamental new insight about the nature of liquid crystalline polymers. By examining the manner in which BCB crosslinking influences the rheology of LCP materials, it may be possible to obtain new information about their mechanisms of flow behavior.

Experimental

The HBA/HNA-co-XTA copolymers were synthesized by adding together stoichiometric amounts of the benzocyclobutene-functionalized crosslinkable monomer XTA and hydroquinone (HQ). The HBA and HNA monomers were kindly provided by Hoechst-Celanese, Inc. The HQ monomer was obtained from Aldrich Chemical Co. The XTA monomer was prepared at the University of Michigan using established procedures (Walker, Markoski, and Moore 1992; Markoski et al., 1993).

The reaction apparatus consisted of a 1000 ml two piece reaction kettle, ChemGlass CG1950-02, three-neck tops CG1960-01, aluminum cased heating mantle, mechanical stirring rod, teflon bearing, short path distillation condensor attached to one neck with a receiving flask, and a thermocouple probe which could reach the surface of reaction media. An additional thermocouple was placed between the kettle and the heating mantle to provide additional information about the temperature of the system.

The dry HBA, HNA, XTA, and HQ monomer powders were weighed to 0.1 mg. 50 mls of acetic anhydride (slightly more than required stoichiometrically) were transferred into the kettle to derivatize the free alcohol groups. 150 ppm (0.015% by weight) of potassium acetate was added as a catalyst.

The reaction mixture was de-gassed by pulling a vacuum to the vapor pressure of the acetic anhydride (evidenced by visible bubbling), then back filling with nitrogen. The stirrer was then turned on and heated up to 120 C for one hour. This step forms acetates from all of the alcohols. After this time, a vacuum was applied for a couple minutes to remove the acetic acid into the condensor. At this point, the mixture gets noticeably thicker as the viscosity increases. The temperature of the kettle is then raised to 300 C.

Once the temperature of the components approaches 300 C, there is an obvious transition from a clear, low viscosity liquid, to a cloudy, high viscosity liquid. The mixture is held at elevated temperature until the laboratory motor can no longer provide sufficient power to stir, typically 30 minutes. During this time a vacuum line is used to remove any unreacted monomer and derivating agents (acetic acid and acetic anhydride). The polymer is removed by quickly disassembling the flask and transferring the hot product onto teflon paper. There is considerable potential for burning the polymer in the reaction flask once the stir bar stops.

The transition temperature from clear to cloudy observed during polymer synthesis evidently corresponds to the onset of liquid crystallinity. This will depend both on molecular weight and on the composition. The final temperature obtained corresponds to the point at which the viscosity of the polymer shuts the motor down. Our observations indicate indicate that the addition of XTA monomer caused a slight decrease in the characteristic temperature of the onset of the liquid crystalline mesophase (from 270 C down to 230 C with the addition of 10% XTA). This is similar to previous observations that the addition of XTA caused a decrease in the transition temperature from the nematic to isotropic phases in lyotropic liquid crystalline PPTA-co-XTA solutions (Jiang et al., 1994).

The 0% HBA/HNA copolymers we synthesized were a dusty brown color, similar to that of the commercial Vectra® materials. With the addition of XTA, the samples developed a more yellow appearance reminiscent of PPTA.

All of the polymers could be ground to a powder at room temperature in a ceramic mortar and pestle. The ground powders were washed in isopropanol for 15 minutes in order to remove any unreacted monomer or residual acetic acid from synthesis. The suspension was then filtered, and dried at 125 C for one hour in a vacuum oven.

The microstructure of the LCPs was examined with Leitz Ortholux II and Nikon Optiphot II Pol optical microscopes and a Linkham LH 1500 hot stage with video overlay system. Quantitative image analysis was facilitated with a Sony CCD video camera connected to a Macintosh Quadra 700 workstation using NIH Image 1.58 software and a RasterOps acquisition board.

All of the polymers were observed to flow and exhibit birefringent nematic textures in cross-polarized light before crosslinking. The size of the polymer particles and fiber fragments was quantified during heating at 10 C/minute in order to reveal the onset of softening and foaming. Swelling studies done on a hot stage, with pentafluorophenol. In this experiment the samples were heated to various temperatures to induce crosslinking, and then reheated to drive off solvent. The dimensions were measured by digital image analysis.

Disks of the polymers for rheological analysis were prepared from 1 gm of the powdered material (25 mm diameter by approximately 3 mm thick) using a Tetrahedron compression mold at 50 psi and 260 C for 15 minutes. Rheological properties were investigated using a Rheometrics RMS-605 rheometer in the parallel plate (torsional shear) geometry for dynamic testing and in the cone-plate geometry for steady testing. Samples were heated by forced convection of dry N_2 gas.

The thermal properties of the polymers were investigated by Differential Scanning Calorimetry (DSC) using a DuPont (at Edwards AFB), and Perkin-Elmer (at Michigan). Best results were obtained when first scanning to 325 C at 20 C/min, then cooling to 50 C, followed by a final scan to 450 C at 10 C/min. The typical DSC sample size was 10 mg. The thermal stability was examined by Thermal Gravimetric Analysis (TGA) using a DuPont (Edwards AFB), Perkin-Elmer (Michigan). The samples (typical size: 25 mg) were scanned in the TGA at 10 C/min in a nitrogen atmosphere.

It was attempted to prepare solutions of the polymers in a 60/40 para-chlorophenol / 1,1,2,2-tetrachloroethane mixture at a 0.1 gm/dl concentration (~20 ml of each solution), as per Bhowmik and Han (1995). However, only limited soluabilty was observed.

The microstructure of the polymer was examined by powder X-ray diffraction using a Scintag 2 kW generator with Cu Ka fixed tube source, Ni filter, theta-theta diffractometer, and tantalum temperature controlled stage with Micristar controller. Scans were taken at room temperature (25 C), 180 C, 260 C, 300 C, and 370 C. Powder scans were also obtained with a Rigaku X-ray generator and theta-theta symmetric reflection diffractometer. The scattering angles ranged from 5 to 75 degrees 2θ.

Fibers of the 5XTA and 10XTA materials were successfully prepared at the University of Michigan using a Bradford University laboratory scale fiber spinning apparatus and take-up line. The sample chamber and piston were 1.27 cm in diameter, with an extrusion ram speed of 0.5 to 2 mm/min. The spinnerette diameter was 250 microns. The typical filament take-up speed was 20 m/min, corresponding to a nominal spin-stretch ratio of 4X. Oriented fibers could also be hand drawn directly from the liquid crystalline melt. in this case the powder was heated on a glass slide on a tube furnace and the fibers were drawn out rapidly with tweezers.

Results

All of the copolymers could be spread into thin films by heating the powder on a microscope in a tube furnace at temperatures of ~350 C. Optical microscopy of the thin films indicated birefringent textures and disclinations, confirming that all of the copolymers would form a thermotropic liquid crystalline mesophase.

Quantitative analysis of the shrinkage during heating was obtained from digital image analysis of powder particles. A plot of the normalized particle area as a function of temperature showed the onset of shrinkage during heating. The temperature corresponding to the onset of shrinkage decreased up to 10XTA, but then increased again for the 20XTA sample.

The DSC traces showed evidence for a Tg near 100 C, which increased with XTA content. There was evidence for a melting endotherm near 280 C in HBA/HNA, but similar transitions could not be unambiguously identified in the copolymers. The onset of the BCB crosslinking reaction was noted as an exotherm near 375 C, as seen in PPTA-co-XTA copolymers. At slightly higher temperatures there were a number of other irregular

peaks which were not reproducible. We associate these peaks with the onset of thermal degradation, as seen in the TGA.

The TGA scans indicated that the onset temperature for thermal degradation decreased with increasing XTA content (Figure 1). However, the amount of residual char seen at elevated temperatures increased systematically with XTA content. The residual char content observed at 900 C increased from 32.7% for HBA/HNA to 39.3% for HBA/HNA-co-20XTA. The decrease in thermal stability with XTA content at temperatures near 400 C was also seen as an increase in the amount of blackening and discoloration of 10 ml glass vials containing equivalent amount of the various copolymers during heating at 400 C in a furnace flushed with nitrogen.

Optical microscopy studies further support the conjecture that it is the onset of the BCB crosslinking reaction that leads to the exotherm above 350 C in the DSC. When the copolymer fibers are heated through this temperature range, the amount of swelling which is observed after exposure to solvent systematically decreases. Figure 2 shows the amount of swelling observed at room temperature for a HBA/HNA-co-10XTA sample as a function of annealing temperature. The plot shows the final diameter divided by the initial diameter of the sample with increasing annealing temperature. Note that the swelling decreases as the annealing temperature increases. Also shown on this plot is the irreversibility of the swelling. The lower curve shows the final diameter of the swollen particle after the temperature was raised to 450 C, driving off the residual solvent. For samples heat treated above 330 C, the swelling is essentially reversible, with the final shape after solvent liberation essentially the same as that of the original sample. However, when the samples which were heat treated at temperatures below 330 C were swollen, an amount of this strain was irreversible, evidently indicating that a permanent crosslinked network had not yet been established.

Rheological Characterization

Initially, temperature sweeps were performed on all of the XTA/HBA/HNA copolymers to complement the DSC results and to establish the processing window within which the polymers would be most easily spun into fibers or molded. Since the crosslinking reaction is kinetically controlled, it is necessary to examine the rate of this reaction to determine the constraints on processability. We chose 5 °C/min as a convenient heating rate for the temperature sweeps and then investigated the kinetics of the crosslinking reaction using isothermal cures. Strain amplitudes were set initially at 1% and lowered during the test to maintain linearity in the rheological material functions. For the crosslinking samples, temperature sweeps were complicated by the fact that expansion of the rheometer tools during heating could not be accomodated by outflow of material (as is usual for polymer melts) and led to a large normal force. Once the normal force exceeded 500 grams, the test was terminated. It was also possible to conduct a manual temperature sweep by raising the top fixture in direct compensation for the thermal expansion.

Figure 3 shows the temperature dependence of the complex viscosity on heating at a rate of 5 C/min. The HBA/HNA (75/25) copolymer which contains no XTA comonomer does not melt until T ~370 C, at which point the complex viscosity falls precipitously. This is in contrast with Vectra A950 data reported in the literature to have a melting point of ~280 C. A possible explanation for this apparent discrepancy might be variations in the HBA/HNA sequence distribution along the chain (Stupp, Fredickson and Leibler). Once the temperature reaches about 390 C, however, the complex viscosity rises, indicating crosslinking even in the absense of XTA comonomer. As pointed out by Economy, crosslinking reactions may be initiated in polyesters by first undergoing the Fries reaction

within the chains. This reaction effectively displaces an oxygen from the ester linkage to a neighboring aromatic ring yielding a hydroxyl substituted aromatic ring adjacent to an ether linkage. The hydroxyl group may then condense with an acidic chain end, leading to crosslinking.

The HBA/HNA-co-1XTA shows no clear melting transition, although the complex viscosity begins a strong decrease at 280 C and shows a viscosity minimum (albeit still large at 2.1) at 318 C. The loss tangent is less than 1 (~.5) for almost the entire temperature sweep, indicating elastic-dominated rheological behavior. The HBA/HNA-co-5XTA copolymer exhibits the same qualitative behavior as the 1XTA sample, with a slightly lower viscosity minimum (~1.1 Poise) which occurs at a lower temperature of 300 C.

The HBA/HNA-co-10XTA copolymer showed quite different behavior in comparison with the previously discussed polymers. The viscosity decreased gradually from the temperature at which the sample is loaded into the rheometer (200 C) and shows a respectably low (1500 poise) viscosity minimum at about 320 C accompanied by a large loss tangent (>2) at the viscosity minimum and followed by a rapid increase in viscosity on continuued heating. Above 350 C, the complex viscosity (* G') becomes larger than 107 dyn/cm^2 at which point transducer compliance effects become significant. The HBA/HNA-co-20XTA polymer showed viscosity increasing with temperature from the moment the temperature sweep was started at 255 C, suggesting that the crosslinking reaction had started before the material had melted all of the crystallites.

While 10XTA (HBA/HNA-co-10%XTA) showed promising rheological results for ease of processibility, it was necessary to monitor the rheological properties versus time for temperatures at or near the viscosity minimum to establish the crosslinking kinetics during conditions required for processing operations such as fiber spinning or extrusion. Three isothermal croslinking reactions were monitored: 255, 275, and 298 C. For each experiment, the 10XTA disk was loaded into the rheometer at the test temperature; the rheometer fixtures had been allowed to reach thermal equilibrium for a period of 30 minutes. Once the samples were loaded, data collection began following a 2 minute delay allowing the temperature of the sample to equilibrate.

The magnitude of G', G", and η* were monitored as a function of time at 255 C for an oscillation frequency of 1 rad/sec and a strain amplitude of 5%. Three time zones are clearly seen. First, the viscosity increases two-fold when first loaded in the rheometer. An explanation may be an increase in molecular weight. Following this initial viscosity increase, there is a period of about 30 minutes over which the rheological properties remain unchanged, following which the moduli begin to increase exponentially, $G(t) = G_o \exp((t-t_o)/t)$. Coinciding with the onset of exponential viscosity increase is the crossing of G' with G", suggesting that the material has undergone a transition through a gel point (Winter). The exponential increase in complex moduli indicates the kinetics of crosslinking are similar to those seen in epoxies (Macosko).

The rate of crosslinking was found to depend strongly on temperature. Shown in Figure 4 is a plot of the complex viscosity monitored during isothermal crosslinking. Several distinct changes occur with increasing the crosslinking temperature. First, of great practical interest is the decrease in incubation time with increasing temperature. This time is reduced from nearly 50 minutes for T = 255 C to less than 20 minutes for T = 298 C. The crosslinking reactions were terminated prior to completion, as the torque signals became too large for the strain amplitude used (which needed to be high enough for meaningful low viscosity data). We analyzed the curves in Figure 4 by fitting the straight portions of the

curves on the log-linear plot to reveal the reaction rate, a, from the fitting equation, $\eta^* = \eta_o^* \exp(a.(t-to))$. A plot of log a vs. 1/T made it possible to estimate the apparent activation energy in the expression $a=a_o \exp(-Ea/RT)$. The apparent activation energy obtained from this analysis is 99 kJ/mol. This number compares reasonably well with values of the heat of reaction of BCB model compounds (ranging from 40 kJ to 208 kJ/mol, typically 82 kJ/mol) as reported by Deeter, et al., (1994).

Given the large incubation time of the HBA/HNA-co-10XTA copolymer at 255 C, we were able to evaluate the steady shear-flow behavior of the uncrosslinked thermotropic melt for a wide range of shear rates. For this test, the cone-plate geometry was used with 25 mm diameter fixtures and a cone angle of 0.04 radians. The shear rate was increased from 0.01 s-1 to approximately 40 s^{-1} at which point fracture of the sample edge became significant. There was strong shear thinning for the entire range of shear rates investigated, with a power-law slope near -0.5, similar to the region I viscosity scaling observed by Onogi and Asada . This behavior is generally consistent with the rheological response of other liquid crystalline polymers, and can be explained by disclination-mediated mechanisms of flow (Wissbrun, Larson and Doi; Kim, Cho, and Chung)

During rheological characterization we found that foaming of the polymer could be observed at high temperatures. This may be related to chain extension due to the liberation of acetic acid end groups. However, no significant mass loss was seen in the TGA. Foaming was also seen in HBA/HNA copolymers without BCB, suggesting that it is not coupled to a reaction involving this moiety.

The thermal and rheological characterization indicate that there is a temperature window for the HBA/HNA-co-XTA polymers within which melt processing can be achieved. A comparison of the various characteristic temperatures are shown in Figure 5 including the glass transition (DSC), softening (OM), minimum viscosity (Rheology), crosslinking exotherm (DSC), and thermal degradation (TGA). The minimum viscosity occurs above the glass transition and softening temperature but below the crosslinking exotherm. Note that the Tg and crosslinking exotherm temperatures both increase slightly with XTA content, while the softening temperature, viscosity minimum, and onset of thermal degradation all decrease with additional XTA.

For the processing of these crosslinkable TLCP's, time and temperature control are critical, because the reaction can continue, leading to curing in the processing equipment. The curable nature of these LCP's provides an additional means for controlling their macroscopic properties. Further work will be necessary to address the impact of this crosslinking on properties of specific interest for a given application. It is also clear that the extent of polymerization at the beginning of processing needs to be examined and controlled more systematically.

X-ray diffraction

The X-ray powder diffraction data are shown in Figure 6. The powder diffraction pattern of the unmodified HBA/HNA copolymer is similar to that observed previously (Guitterez et al., 1983, Cheng et al., 1989; Flores, et al., 1993). There is a strong peak a 4.6 A, a medium peak at 6.5 A, a medium peak at 3.3 A, and a weak peak at 2.1 A. Previous work has associated these peaks with the (110), (100), (211/210), and meridional reflections of an orthorhombic unit cell similar to that of the HBA homopolymer (a=7.62 A, b=5.70 A, c=12.56 A) (Kaito, et al., 1990). There were also three sharp peaks located at d-spacings of 2.82, 1.99, and 1.63 A. The fiber X-ray patterns confirmed that these

correspond to crystalline aggregates which evidently associated with the catalyst or a complex between the catalyst one of the added comonomers, since their intensity increases systematically with XTA content. We have found that the peaks can be removed by washing the ground powder for 24 hours in distilled water.

With the addition of the XTA monomer, there are a number of systematic changes to the copolymer peaks in the X-ray diffraction patterns.:

1. The strong (110) peak near 4.6 A systematically weakens in intensity and broadens in width with increasing XTA content. The peak position moves to lower angle for 1 XTA, but then moves outward as the XTA content continues to increase.

2. The medium (100) peak near 6.6 A remains similar in intensity and width with increasing XTA content, but is missing altogther in the HBA/HNA-co-10XTA material. The peak position initially moves to a higher angle, but then moves inward as the XTA content increases.

3. The (211/210) peak near 3.3 A first sharpens and moves out to higher angle, then moves inward and fades in intensity with XTA content.

4. The peak near 2.08 A remains in nearly the same position with increasing XTA content, although it decreases in intensity and broadens somewhat.

5. The HBA/HNA-co-1XTA sample shows a new peak at 3.83 A that is not present in any of the other samples. The intensity and width of this peak is similar to that of the other reflections observed in HBA/HNA.

Summary of d-spacings from X-ray powder data:

XTA Content	Peak #1 medium stable broad	Peak #2 strong decays widens	Peak #3 medium only in 1XTA	Peak #4 medium decays broad	Peak #5 weak decays broad
0	6.647	4.567	--	3.308	2.078
1	6.458	4.610	3.829	3.159	2.081
5	6.539	4.549	--	3.268	2.07
10	--	4.547	--	3.271	2.07
20	6.784	--	--	3.274	2.07

Figure 7 shows fiber x-ray diffraction patterns from oriented filaments of HBA/HNA-co-5XTA and HBA/HNA-co-10XTA copolymers, (1) as-spun from the melt state and (2) after heat treatment in a nitrogen atmosphere for two minutes. These patterns confirm that oriented fibers of the copolymers can be obtained by melt processing, and that their structural perfection can be improved by post-spinning annealing treatments. Additional characterization of the microstructure and mechanical properties of these fibers are currently underway.

Hypotheses

Our TGA data indicates that the addition of the BCB functional gropus causes a decrease in the onset of thermal degradation of HBA/HNA thermotropic copolyesters. This reduction in thermal stability (from 505 C to 450 C with 20%XTA) is more dramatic than

was seen previously in the case of aromatic polyamides (from 530 C to 475 C for 100% XTA in PPTA) (Deeter, 1994). This might be due to the fact that a number of other types of reactions which are known to occur in polyesters such as transesterifcation and Fries reaction, and the BCB crosslinking may be accelerating or modifying these reactions. These reactions may also be influenced by presence of oxygen. These results suggest that BCB moieties will be useful for improving flame resistance only in situations where the increased resistance to melt flow is more important than the initial stages of thermal degradation.

There are a number of different reactions which can take place in the 300-400 C range in these thermotropic copolyesters, including melting, crystallization, transesterification, chain extension, Fries rearrangements, foaming, thermal degradation, textural relaxations, and chain scission. By adding BCB functional groups via the XTA comonomer, it appears that the solid-state chemistry of of these materials has become even more complicated. Further work will be necessary to better establish the influence of these various phenomena on macroscopic properties and the long-term stability and performacne of these materials.

To further understand the behavior of this system, it will be critical to establish the kinetics associated with these different phenomena, and relate this information to macroscopic observations of performance. In this manner, it will be possible to determine the utility of these materials in lightweight structural components. It will also be possible to identify more clearly identify the limitations, and hopefully motivate additional efforts in monomer and polymer design, synthesis, and characterization. In any case, the ability to manipulate and control the chemical crosslinking of thermotropic copolyesters provides for a number of new opportunites for future liquid crystalline polymer materials research.

Conclusions

1. HBA/HNA-co-XTA copolymers with systematic variations in XTA content can be synthesized and characterized.

2. There is a processing window near 280 C above the softening temperature (200-250 C) but below the crosslinking reaction (350 C) within which the polymers exhibit a viscosity minimum.

3. When held at elevated temperatures, the polymers exhibit an increase in viscosity with time that seems consistent with the formation of crosslinks due to the reaction of the BCB.

4. The addition of XTA causes a decrease of thermal degradation onset temperature, but an increase in the char yield at 900 C. These char yields are above 35% for all of the materials tested.

5. The addition of XTA units to the HBA/HNA polymer backbone makes it possible to explore the microstructure of the polymers in more detail. Further studies are necessary to refine our model of the thermotropic liquid crystal polymer solid-state.

6. The feasiblity of melt spinning fibers from the HBA/HNA-co-XTA copolymers has been confirmed. Further work is necessary to optimize processing conditions and properties. The melt processing of these thermally crosslinkable copolyesters will require careful temperature control in order to avoid premature curing in the spinning chamber or extruder.

Future

Additional information about the mechanisms of the BCB crosslinking reaction could be obtained by spectroscopic techniques such as Fourier Transform Infrared Spectroscopy (FTIR) or solid-state carbon-13 Nuclear Magnetic Resonance (NMR). Of particular interest would be determining whether or not there is chain scission which accompanys the BCB reaction. Without such information, understanding the origin of the changes in macroscopic properties near 300 C will be difficult because of the many different possibilities for both chemical and physical structure. These effects include: crystallization, melting, crosslinking, degradation, foaming, molecular weight extension, and transesterfication. Some of these processes may be interdependent. In other words, crystallization can be facilitated by transesterification, and the transesterification reaction itself can be driven by crystallization (Lenz, CIR). It has also been suggested that certain processes of degradation and foaming can be induced by particular processing schemes (Economy, 1994).

Additional efforts will be placed on monitoring the entire crosslinking reaction. It is expected that plots of complex viscosity versus time will yield curves of sigmoidal shape, the rate of reaction reaching a maximum at some point. With this data, we will be able to analyze the crosslinking kinetics using methods analogous to chemorheological epoxy studies.

References

"Synthesis of Thermotropic Polyesters", U. S. Patent Nos. 4,161,470, 4,429,105.

K. F. Wissbrun, J. Rheology, 25(6), 619-662, (1981).

J. Blackwell and G. A. Gutierre, Polymer, 23, 671, (1982).

K. F. Wissbrun, Faraday Disc. Chem. Soc., 79, 161-173, (1985).

S. Z. D. Cheng, J. J. Janimak, A. Zhang, and Z. Zhou, Macromolecules, 22, 4240-4246, (1989).

D. S. Kalika, D. W. Giles, and M. M. Denn, J. Rheol., 34(2), (1990).

A. Kaito, M. Kyotani, and K. Nakayama, Macromolecules, 23, 1035-1040, (1990).

G. H. Frederickson and L. Leibler, Macromolecules, 23, 531-539, (1990).

R. G. Larson and M. Doi, J. Rheology, 35(4), 539-563, (1991).

C. L. Jackson and M. T. Shaw, International Materials Reviews, 36(5), 165-186, (1991).

T. Pakula and R. Zentel, Makromol. Chem., 2401-2410, (1991).

W. Lee and A. T. Dibenedetto, Polymer Eng. and Sci., 32(6), 400-408, (1992).

K. A. Walker; L. J. Markoski; and J. S. Moore, Synthesis, 1265-68, (1992).

W. Lee, A. T. Dibenedetto, J. M. Gromek, M. R. Nobile, and D. Acierno, Polymer Eng. and Sci., 33(3), 156-165, (1993).

A. Flores, F. Ania, and F. J. Balta-Callega, Polymer, 34(14), 2915-2920, (1993).

L. J. Markoski, K. A. Walker, G. A. Deeter, G. E. Spilman, D. C. Martin, and J. S. Moore, J. S. Chem. Mater., 5, 248 (1993).

K. M. Kim, H. Cho, and I. J. Chung, J. Rheology, 38, 1271-1283, (1994).

J. Economy, D. Frich, K. Goranov, and J. C. Lim, Polymer Mat. Sci. Eng., (1994).

M. J. Marks, J. S. Erskine, and D. A. McCrery, Macromolecules, 27, 4114-4126, (1994).

M. J. Marks and J. K. Sekingesr, Macromolecules, 27, 4106-4113, (1994).

K. L. Walker, L. J. Markoski, J. S. Moore, and D. C. Martin, U. S. Patent No. 5,334,752, (1994).

C. W. Potter, J. C. Lim, G. Serpe, and J. Economy, in Progress in Pacific Polymer Science 3, K. P. Giggino, ed., 271-281, (1994).

S. B. Warner and J. Lee, J. Poly. Sci.: B: Poly. Phys. Ed., 32, 1759-1769, (1994).

C. M. McCullagh, J. Blackwell, and A. M. Jamieson, Macromolecules, 27, 2996-3001, (1994).

G. A. Deeter, D. Venkataraman, J. W. Kampf, and J. S. Moore, Macromolecules, 27, 2647-2657, (1994).

G. W. Meyer, S. J. Pak, Y. J. Lee, and J. E. McGrath, Polymer, 36(11), 2303, (1995).
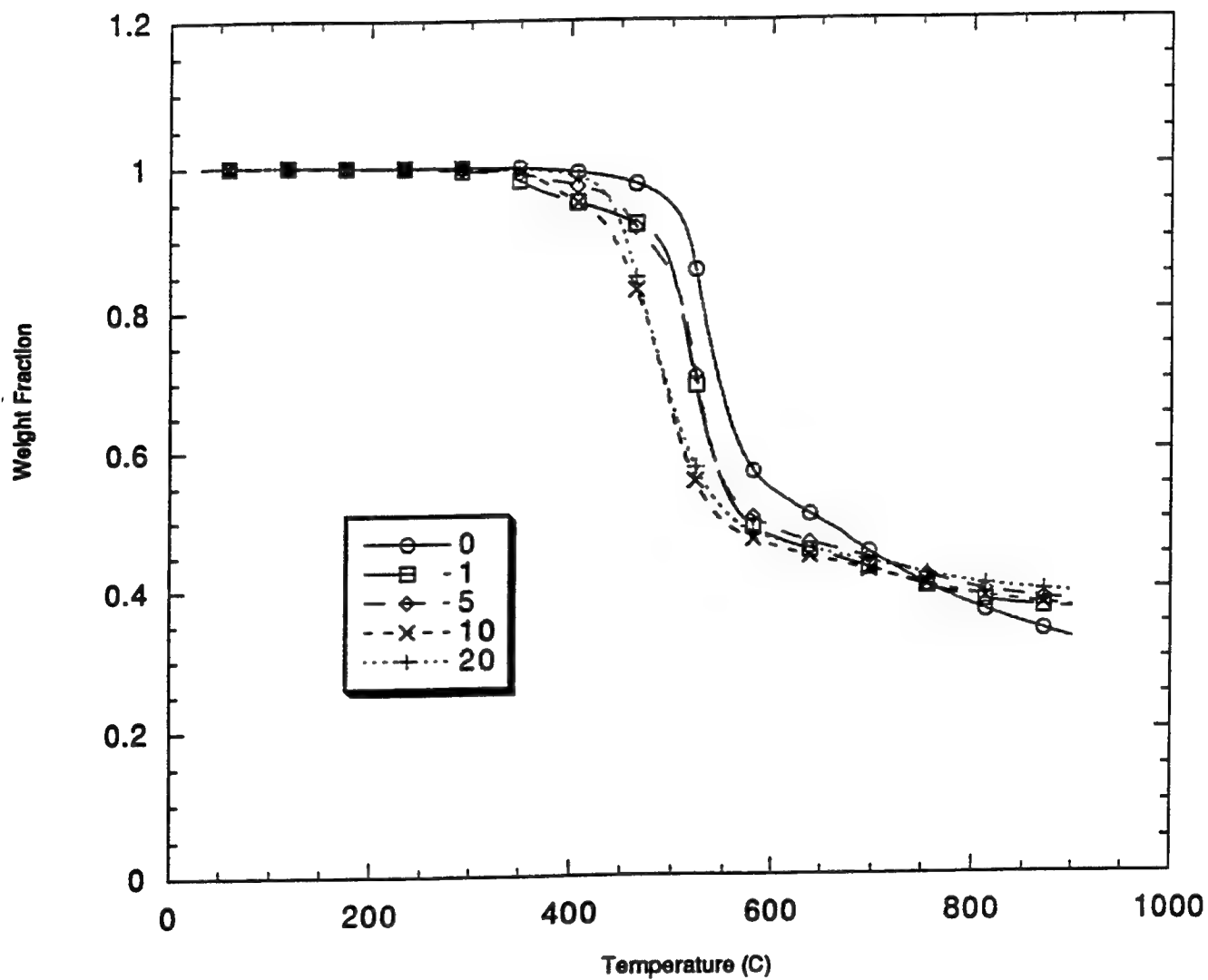
HBA/HNA-co-XTA   TGA

Figure 1: TGA traces of the HBA/HNA-co-XTA copolymers as a function of XTA content. The XTA-containing copolymers are less stable near 400 C, but show improved char formation above 700 C.
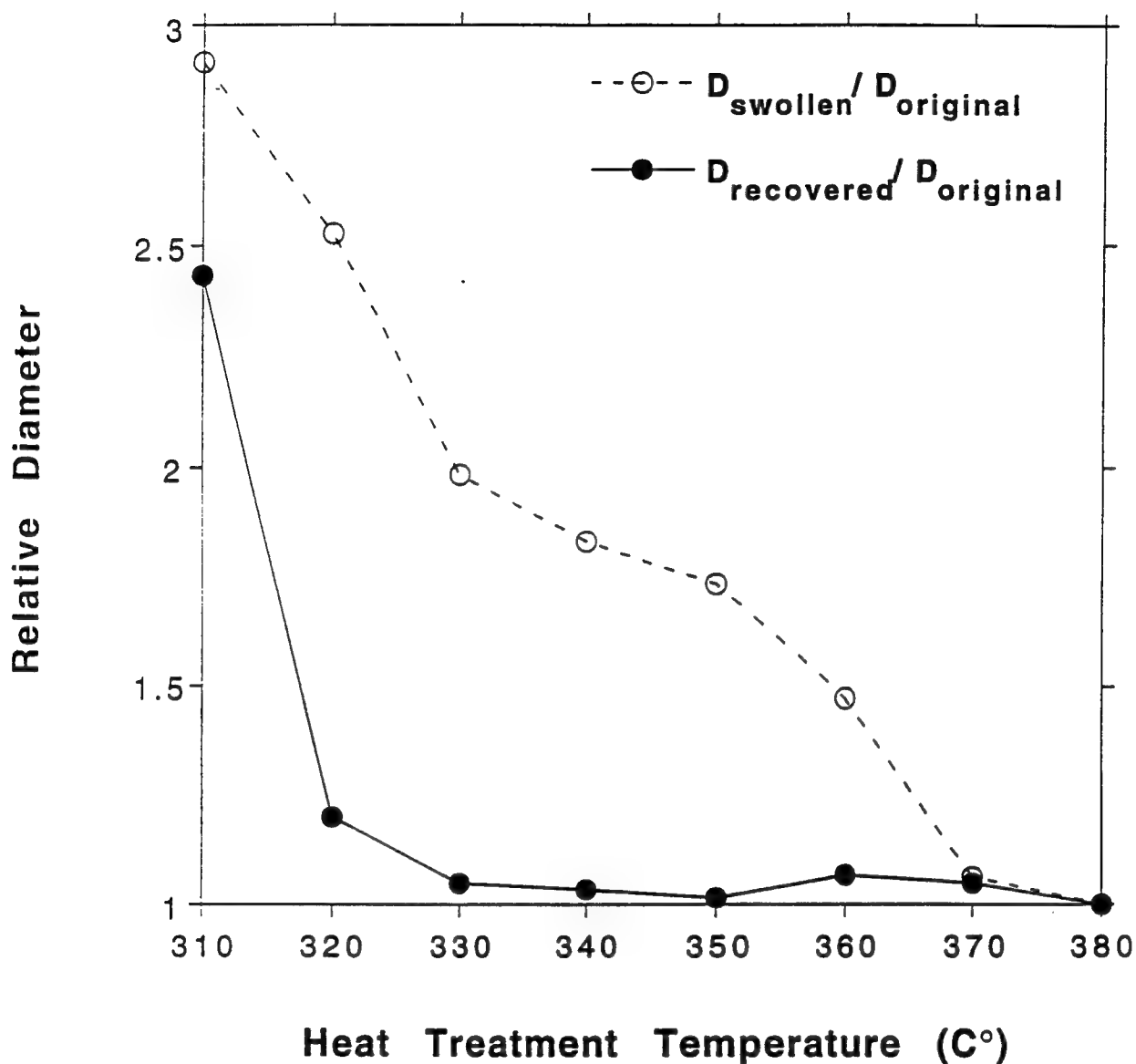
Figure 2: Change of the diameter of the HBA/HNA-co-10XTA copolymer fibers during swelling in pentachlorophenol after heat treatments at various temperatures. As the heat treatment temperature increases through the BCB reaction temperature, the amount of swelling systematically decreases. After the sample is reheated, the solvent is driven off and the particle recovers its original shape (above 330 C). For the most lightly crosslinked materials (heat treatment temperatures below 330 C), the swelling is irreversible.
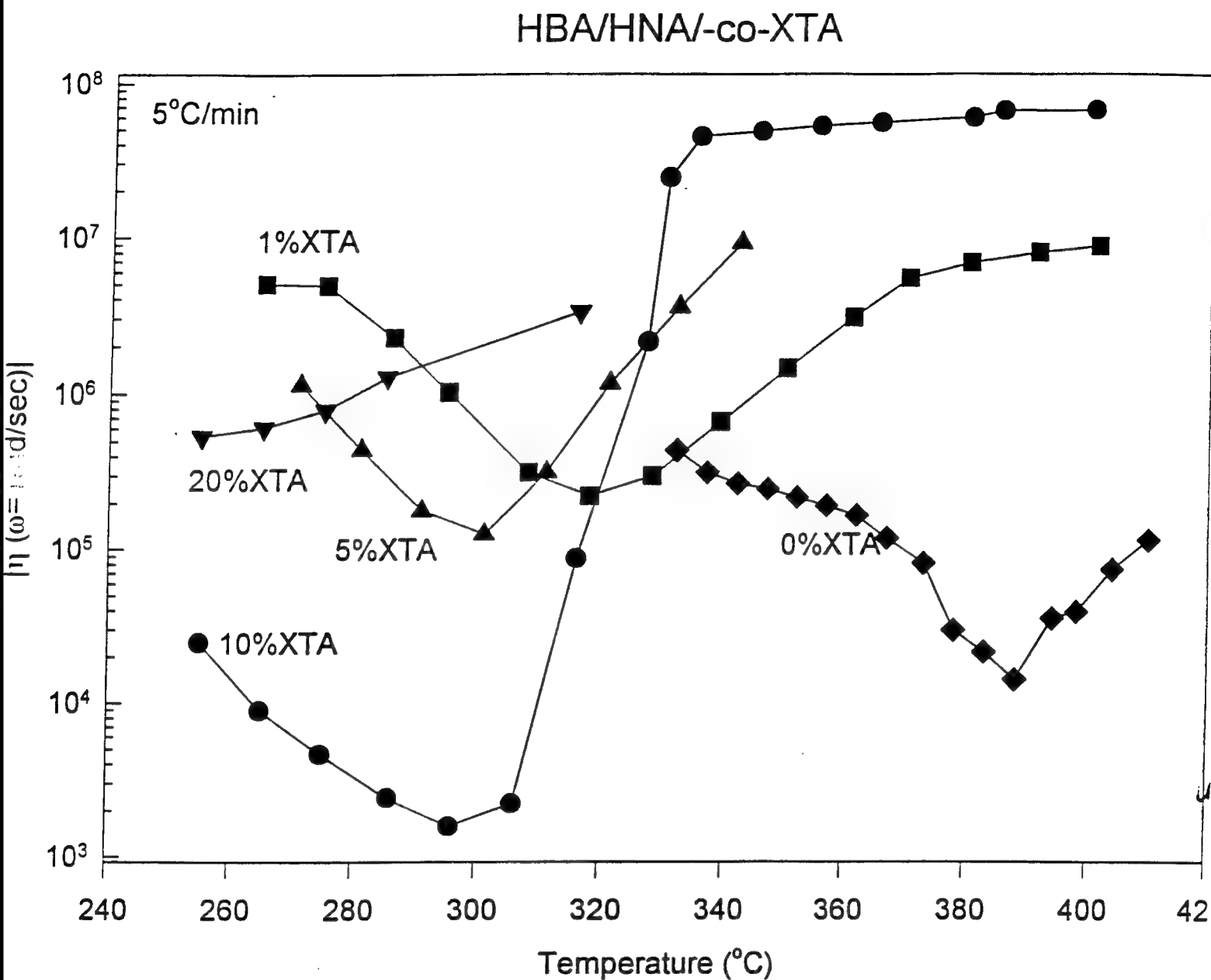
HBA/HNA/-co-XTA

Figure 3: Complex viscosity of the HBA/HNA-co-XTA copolymers as a function of XTA content and temperature. The minimum viscosity occurs at lower temperature with increasing XTA content. For the HBA/HNA-co-10XTA copolymer, there is a 5 order of magnitude increase in the complex viscosity near 320 C as the sample crosslinks.
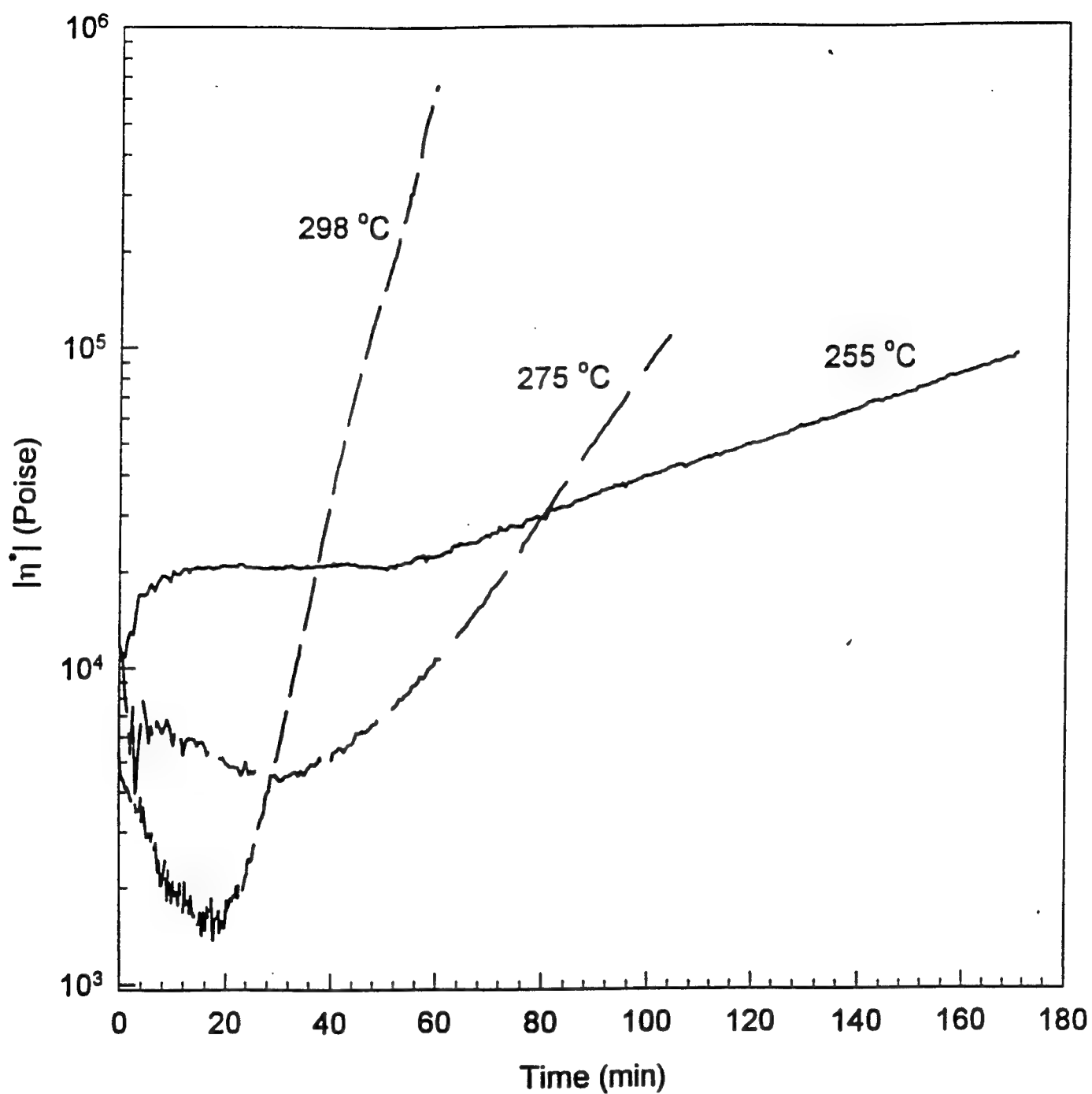
Figure 4: Plot of the complex viscosity of the HBA/HNA-co-10XTA copolymer as a function of time at various temperatures. As the temperature increases, the rate of crosslinking increases and the incubation time decreases.
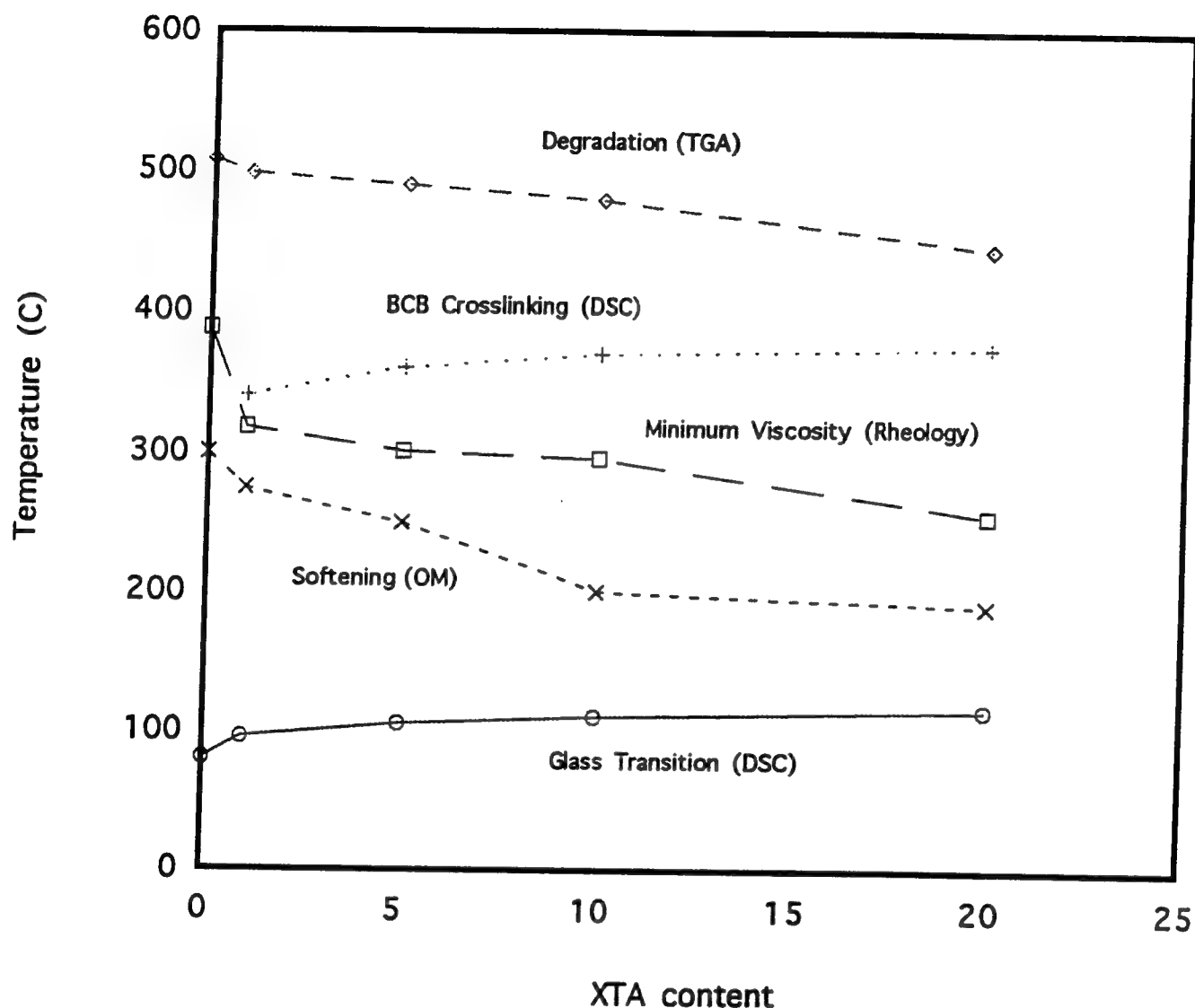
Figure 5: Characteristic temperatures of the HBA/HNA-co-XTA copolymers as a function of XTA content. The glass transition (DSC, Perkin-Elmer and DuPont) increases slightly with XTA content. The softening temperature was determined by hot stage optical microscopy (Linkham, Leitz, Nikon), and decreases with XTA content. The minimum viscosity data was obtained by rheology (Rheometrics), and decreases with XTA content. The crosslinking temperature was determined by DSC, and increases with XTA content. The degradation temperature was determined by TGA (Perkin-Elmer, DuPont) and decreases with XTA content.

The experimental data indicate that there is a processing window for the HBA/HNA-co-XTA copolymers roughly indicated by the minimum viscosity temperature. This processing window is above the softening temperatures but below the crosslinking and degradation temperatures.
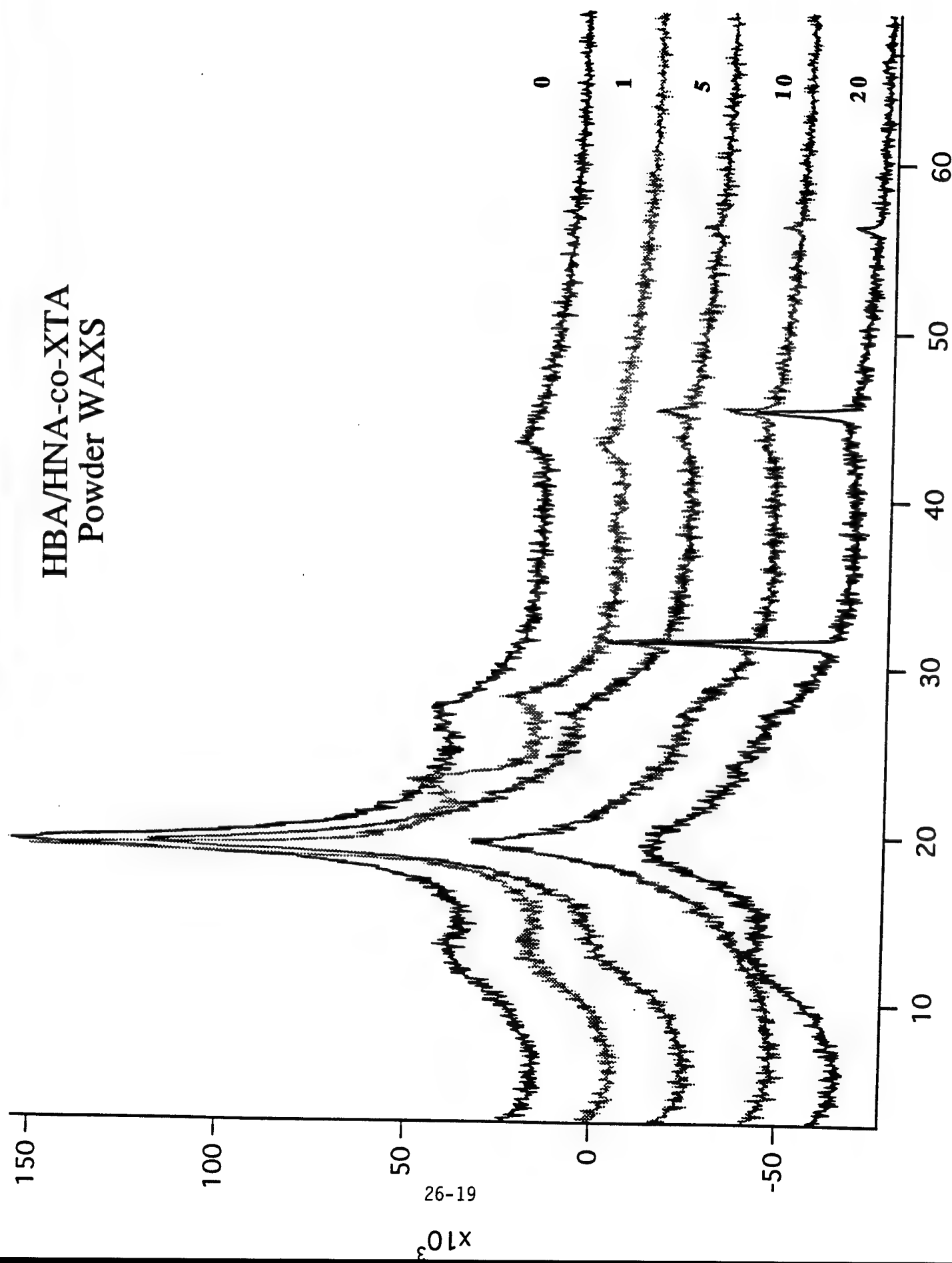
Figure 6: Powder WAXS patterns of HBA/HNA-co-XTA as a function of XTA content.

As Spun

Heat Treated
200 C

5%XTA
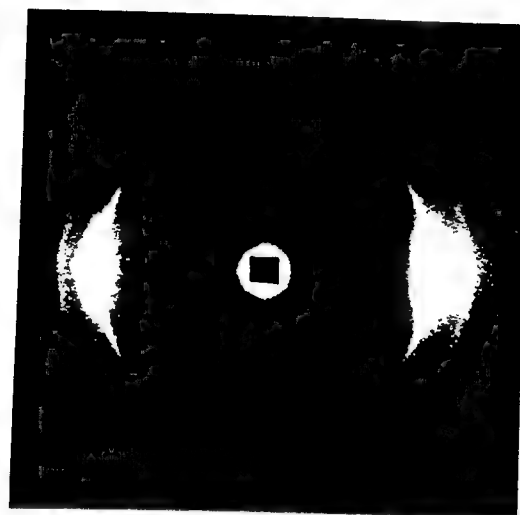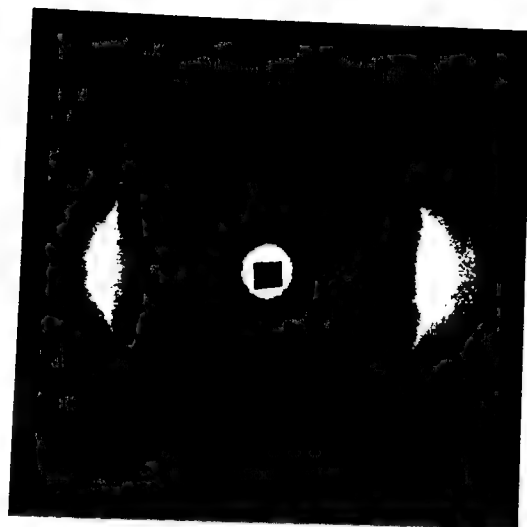
10%XTA
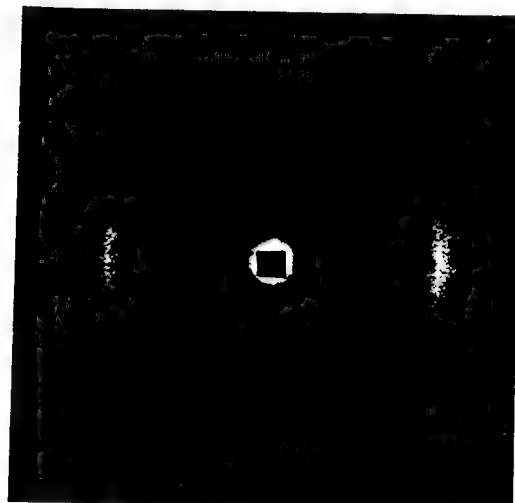
Figure 7: Fiber WAXS patterns of HBA/HNA-co-XTA copolymer fibers as-melt spun and after heat treatment at 200 C.

SYNTHESIS OF PRECURSORS TO ENERGETIC CATIONS

G. David Mendenhall
Professor
Department of Chemistry

Michigan Technological University
1500 Townsend Dr.
Houghton, MI 49931

Final Report for:
Summer Faculty Research Program
Phillips Laboratory

SYNTHESIS OF PRECURSORS TO ENERGETIC CATIONS

G. David Mendenhall
Professor
Department of Chemistry

Michigan Technological University

Abstract

A number of promising intermediates for the synthesis of energetic cations have been synthesized from procedures in the literature or modifications of them, including nitroform and its salts, dichlorodinitromethane, silver azide, and others. A novel potential intermediate, nitrosotrinitromethane, appeared to decompose at $0^{\circ}C$ to give nitric oxide, carbon dioxide, and other gases. Tetra-n-butylammonium azide was shown to be a promising reagent for extremely fast nucleophilic introduction of the azide group.

SYNTHESIS OF PRECURSORS TO ENERGETIC CATIONS

G. David Mendenhall

Introduction

The Air Force has an interest in the development of rocket propellants of increased energy density. The oxidizer portion of propellants typically consists of an anion in a highly oxidized state such as nitrate or perchlorate. The counterion is characteristically ammonium or a metal ion that does not contribute to oxidizing ability. The pairing of a highly oxidized cation with a highly oxidized anion is a conceptual approach to enhance the oxidizing component of the fuel. The existence of $(N_3)_3C^+SbCl_6^-$ as a crystalline solid provides a stimulus to explore the synthesis and properties of related cationic species.

The efforts during the summer involved synthesis and characterization of a number of potential precursors to the novel cations $(N_3)_n(O_2N)_{3-n}C^+$ and $[(O_2N)_3C]_2NO^+$. Since it is difficult to predict what approach would succeed, we have begun by preparing (a) a number of precursors of the type $CX_n(NO_2)_{4-n}$, which may be converted to azidonitromethanes with AgN_3 or a soluble azide, and (b) dinitrodiazomethane, which may react with halogen to form dinitrodihalomethanes, or with HN_3 or HX to form substituted dinitromethanes.
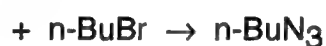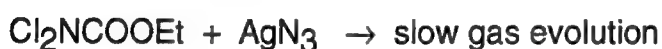
Results and discussion

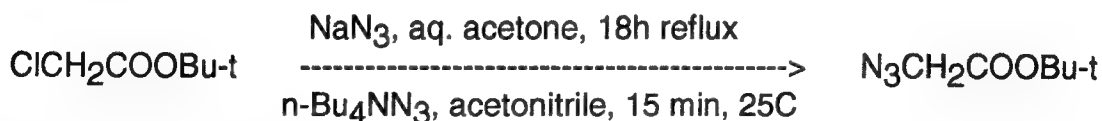Some of the synthetic reactions that were carried out are as follows:

$$C(NO_2)_4 + K_4Fe(CN)_6 \rightarrow NO_2^- + KC(NO_2)_3(s) \rightarrow HC(NO_2)_3 \quad [1]$$

$$KC(NO_2)_3 + Ph_4SbBr \rightarrow Ph_4SbC(NO_2)_3 \quad [2]$$

$$+ Ph_4AsCl \rightarrow Ph_4AsC(NO_2)_3$$

$$+ AgNO_3 \rightarrow AgC(NO_2)_3$$

$$AgNO_3 + NaN_3 \rightarrow AgN_3$$

$$n\text{-}Bu_4NOMe + HN_3 \rightarrow n\text{-}Bu_4NN_3 + MeOH \qquad [3]$$

$$n\text{-}Bu_4NBr + AgN_3 \rightarrow n\text{-}Bu_4NN_3$$

$$ClCH_2COCl + t\text{-}BuOH \rightarrow ClCH_2COOBu\text{-}t \qquad [4]$$

$$ClCH_2COOBu\text{-}t + n\text{-}Bu_4NN_3 \rightarrow N_3CH_2COOBu\text{-}t \qquad [5]$$
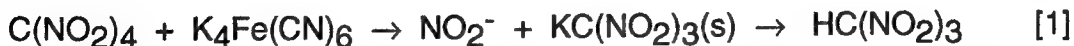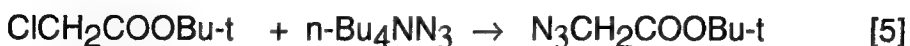
In addition, some reactions that were carried out qualitatively (GC, GCMS or visual observations) in preliminary work were as follows:

$$CCl_2(NO_2)_2 + AgN_3 \rightarrow products$$

$$Me_3Si\text{-}N_3 + Me_4NCl \rightarrow no\ reaction$$

$$Cl_2NCOOEt + AgN_3 \rightarrow slow\ gas\ evolution$$

$$n\text{-}Bu_4NN_3 + C(NO_2)_4 \rightarrow vigorous\ gas\ evolution$$

$$+ n\text{-}BuBr \rightarrow n\text{-}BuN_3$$

$$+ [CH_2Br_2, CHBr_3, CBr_4, Br_2CHCN] \rightarrow products$$

Since tetra-n-butylammonium azide was found to liquify in the presence of small amounts of water or organic solvents, the compound can furnish a highly nucleophilic medium containing azide ion at concentrations of 2M or higher. Preliminary indications are that this expectation is realized. For instance, the procedure in Organic Synthesis to prepare tert-butyl azidoacetate calls for a lengthy reflux of the chloroacetate with sodium azide, whereas in our hands tetra-n-butylammonium azide was found to react exothermically with the chloracetate at room temperature.

$$ClCH_2COOBu\text{-}t \xrightarrow[\text{n-Bu}_4\text{NN}_3,\ \text{acetonitrile, 15 min, 25C}]{\text{NaN}_3,\ \text{aq. acetone, 18h reflux}} N_3CH_2COOBu\text{-}t$$
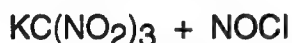
Nitration of trichloroethylene

Burrows and Hunter[6] reported a marvellously simple approach to a potential intermediate for the goals of this program:

27-4

$$CCl_2=CHCl \xrightarrow[\text{reflux, 3h}]{HNO_3} CCl_2(NO_2)_2 + \text{other products}$$
$$17\%$$
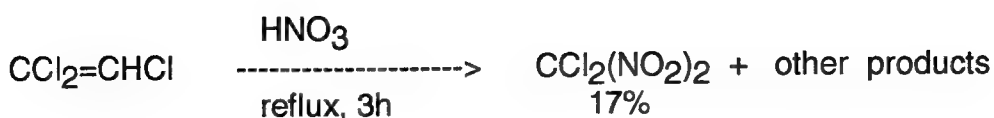
Dichlorodinitromethane was isolated in 90% purity by fractional distillation of the steam-distillate from this reaction. Burrows and Hunter noted two minor products from the reaction and identified them as chloropicrin (CCl_3NO_2) and unreacted trichloroethylene. In our hands there were two side products in the GC trace of our redistilled material. The side products were concentrated in the higher-boiling fractions, and neither retention time corresponded to that of authentic chloropicrin or starting trichloroethylene. Although the mass spectrum of the larger unknown resembled that of tetrachloroethylene, the GC retention of an authentic sample was much shorter.

Attempted synthesis of nitrosotrinitromethane (nitrosyl nitroformate)

Attempts to synthesize this unknown compound were carried out with the following approaches:

$$KC(NO_2)_3 + NOCl$$
$$Ph_4AsC(NO_2)_3 + NOBF_4 \quad \xrightarrow{\text{-------//---------}} \quad O=N-C(NO_2)_3$$
$$KC(NO_2)_3 + NOBF_4$$

In all cases brown gases, presumably NO_2, appeared above the surface of the reaction mixtures that were carried out in air. The GC traces of filtered recation mixtures revealed no compounds with retention times that could be ascribed to nitrosyl nitroformate, and concentration under reduced pressure led only to traces of nonvolatile residue. In some reactions small amounts of nitroform were detected chromatographically. This could have arisen from H-abstraction from solvent by the radical $\bullet C(NO_2)_3$ or from traces of water in the starting materials.

The heterogeneous reaction of potassium nitroformate with nitrosyl fluoroborate in MeCN was also carried out in a flask attached to a vacuum line. The volume of the system was measured beforehand, so that the amount of gas evolved could be converted directly to equivalents. The data, shown in Figure 1, are interesting in that they hint of a reaction proceeding in two stages.

Mass spectral analysis of the evolved gases from this reaction was consistent with the predominence of NO (m/e 30), CO_2 (m/e 44), and nitrogen or CO (m/e 28). A pale brown color in the gas sample deepened when air was admitted to the flask, consistent with the conversion of NO to NO_2.

The results of this experiment were consistent with a successful synthesis of an unstable compound. In a reaction designed to test the appropriateness of our approach, we were gratified by the successful preparation of tetranitromethane under identical conditions:

$$KC(NO_2)_3 \ + \ NO_2BF_4 \quad \xrightarrow[\text{0}^{\circ}\text{C}]{\text{MeCN, 2 min}} \quad \underset{67\%}{C(NO_2)_4}$$

However, we were surprised at the complete failure of the approach with more soluble ionic compounds (more suited for low-temperature reactions) in a different solvent:

$$Ph_4AsC(NO_2)_3 \ + \ NO_2PF_6 \quad \xrightarrow[\text{0-25C}]{\text{CH}_2\text{Cl}_2, \ 30 \text{ min}} \xcancel{\quad} \quad C(NO_2)_4$$
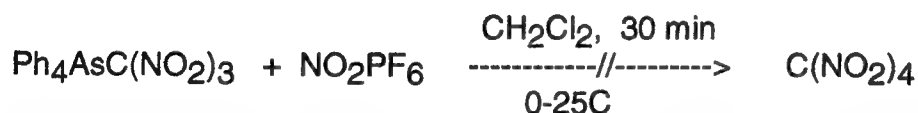
There are several possible explanations for this result, but the reaction was not studied further.

Conclusions

A number of promising intermediates for the synthesis of energetic cations have been synthesized. Tetra-n-butylammonium azide appears to be superior to existing reagents for the nucleophilic displacement reactions to introduce the azide group. The compound nitrosotrinitromethane does not appear to be stable at 0^oC.

References

1. F.D. Chattaway and J.M Harrison, J. Chem. Soc. 1916, p. 171.

2. V.I. Erashko and S.A. Shevelev, Izv. Akad. Nauk SSSR, Ser. Khim. 1985, 439-42.

3. V. Gutmann, G. Hampel, and O. Leitmann, Monatshefte fur Chemie; **95**, 1034 (1964).

4. R.H. Baker, Org. Synth. Coll. Vol. III, p. 144.

5. A.T. Moore and H.N. Rydon, Org. Synth. Coll. Vol. V, p. 586.

6. R.B. Burrows and L. Hunter, J. Chem. Soc. 1932, p. 357-60.
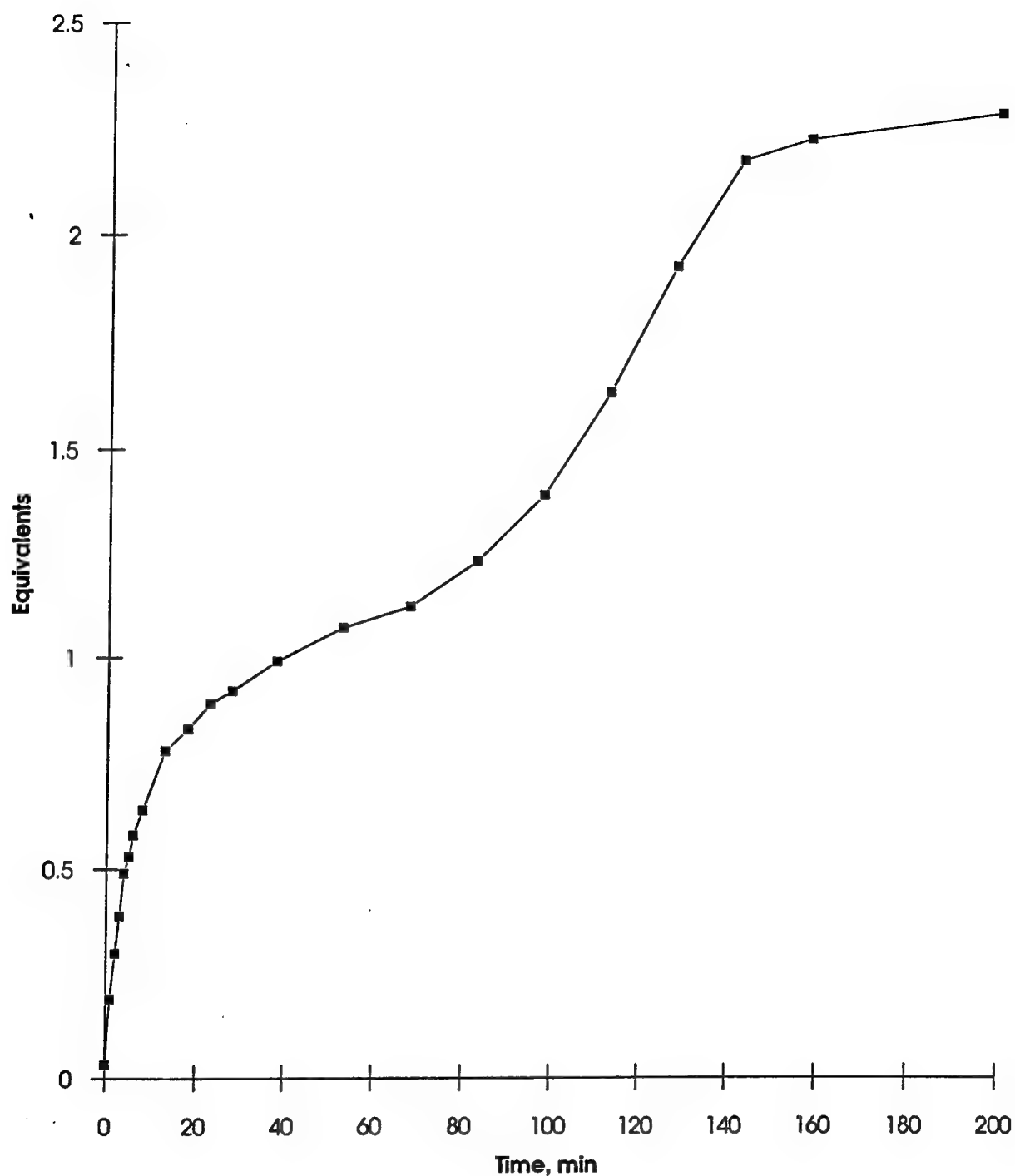
Figure 1. Gas evolution from reaction of 1.5 mol $NOBF_4$ and 1.5 mol $KC(NO_2)_3$ in 1.0mL acetonitrile at 0^oC, corrected for vapor pressure of solvent. The point shown at 200min was recorded after allowing reaction to warm overnight followed by recooling to 0^oC.

Theory of the Equatorial Spread F: Linear and Nonlinear Dynamics and Wavepackets

Stefano Migliuolo

Research Physicist

Research Laboratory of Electronics

Massachusetts Institute of Technology

77 Massachusetts Avenue

Cambridge, MA 02139

Theory of the Equatorial Spread F: Linear and Nonlinear Dynamics and Wavepackets

Stefano Migliuolo

Research Physicist

Research Laboratory of Electronics

Massachusetts Institute of Technology

Abstract

The linear stability and eventual nonlinear saturation of plasma depletions in the high altitude ionosphere, the F-layer, are modelled via two–fluid equations for a plasma consisting of electrons and a single species of ions. Major emphasis is given to the non–local properties of the perturbations, permitting the description of relatively long wavelength modes ($k_\perp L \sim 1$, where L is the plasma density gradient scale length and $k_\perp$ the wavevector perpendicular to the magnetic field), a regime not accessible under the often used "local approximation". In examining the linear theory, attention is given to effects due to parallel dynamics, that can give a strong stabilizing contribution for disturbances with perpendicular wavelengths longer than 1 km (i.e., where the power spectrum of Spread F is maximum). The condition for onset (marginal stability) is derived analytically for the ionosphere, including a height–dependent recombination rate. The variation of the eigenfunction with height is addressed using two model equilibria: an exponentially stratified atmosphere and one with local maxima in the density gradient. A nonlinear model, comprising both quasilinear and three–wave coupling, is used to illustrate the dynamics of the depletions and estimates for the saturated amplitudes are given.

Theory of the Equatorial Spread F: Linear and Nonlinear Dynamics and Wavepackets

Stefano Migliuolo

1. Introduction

Since the earliest measurements obtained from ionosondes, the ionosphere has been know to be both stratified and dynamically active. A class of irregularities, i.e. local changes in the ambient plasma density, that has received considerable attention is the so–called "spread F". These are generally depletions (bubbles) detected from ground radar (backscatter) and satellites (scintillation), with scales ranging from a few meters to several kilometers (cf. *Basu et al.*, [1978]), occurring preferentially during nighttime, starting in the lower portion (bottomside) of the F-layer (200–300 km altitude) and rising to the topside, sometimes up to 900 km altitude (cf. *Kelley*, [1989]). These irregularities are believed to arise from plasma instabilities engendered by a change in local properties of the plasma. There exist a number of reviews on this topic, in particular we mention those by *Fejer and Kelley* [1980], and *Ossakow* [1979]–[1981].

Considerable work has been carried out to determine the linear stability properties of the plasma in the nighttime equatorial sector, with the objective to identify both the nature of the governing instability as well as critical values of plasma parameters. The most promising candidate for the generation of these irregularities is a plasma instability, which can be termed "generalized $\mathbf{E} \times \mathbf{B}$" and that arises due to the natural separation of charge in a plasma that is subject to a combination of electric field or gravity (or both) and a density gradient. When gravity is dominant, it is known as the collisional Rayleigh–Taylor instability (e.g., *Dungey*, [1956], *Balsley et al.* [1972], *Haerendel* [1974]). Otherwise, it is known as the $\mathbf{E} \times \mathbf{B}$ or gradient-n instability (*Simon* [1963], *Hoh* [1963]). Examples of the complementary roles played by gravity and electric fields can be found, e.g., in the papers by *Perkins et al.* [1973] and *Huba et al.* [1983]. A comprehensive list of publications is beyond the scope of this paper, but we mention *Hudson and Kennel* [1975a,b] who included thermal effects as well as coupling to the Alfvén wave, *Guzdar et al.* [1982] and *Satyanarayana et al.* [1984] who investigated the influence of velocity shear on the basic Rayleigh–Taylor instability, *Basu and Coppi* [1983] and *Glassman and Sperling* [1983] who initially looked at effects of finite parallel wavelength and investigated the coupling to

electromagnetics and parallel dynamics (note that *Basu and Coppi* [1983]–[1984] were the first to look at this problem from the point of view of a two–dimensional magnetic field).

Building on early work by *Rognlien and Weinstock* [1974] that dealt with the electrojet, *Chaturvedi and Ossakow* [1977] laid the foundation for present nonlinear models that model irregularities in the ionosphere, both equatorial and auroral. The primary nonlinearity arises in the equation of conservation of particle number (the "continuity equation") and is applicable to a number of interchange–like modes, in addition to the Rayleigh–Taylor instability: the gradient drift instability *Chaturvedi and Ossakow* [1979a] and the current–convective instability [1979b] (this latter mode is driven unstable by an equilibrium current parallel to the ambient magnetic field and is pertinent to the physics of, e.g., auroral arcs). Of note are also papers by *Ott* [1978] who gave an analytic estimate of the size and vertical velocity of the bubble and *Hudson* [1978] who included ion inertia. The current emphasis of nonlinear work is in the use of self–consistent numerical codes that give a rather complete description of the time evolution of plasma irregularities in two dimensions, cf. *Ossakow et al.* [1979], *Keskinen et al.* [1980], [1981], *Hassam et al.* [1986].

The large majority of the analytic work, linear and nonlinear, has been carried out in the so–called local approximation, where the wavelength components perpendicular to the ambient magnetic field are assumed to be much shorter than any equilibrium scale length. This is certainly justified in cases where, e.g. the density gradient scale length is of the order of 5–20 km, while the perturbations of interest have sub–kilometer wavelengths. However, this approximation becomes increasingly irrelevant as one attempts to look at long wavelength irregularities (for which the power spectrum is maximum) or in regions of enhanced plasma density gradients. Thus, it is of interest to re–visit the theory without making this approximation to see whether insight may be gleaned in the dynamics of per-turbations with wavelengths comparable to the equilibrium scale lengths.We will include parallel dynamics and recover the result of *Basu and Coppi* [1983]–[1984]): a substantial linear stabilization arises. We extend their analysis by showing that parallel dynamics can only stabilize (never destabilize) these long wavelength modes in the collisional plasma of the F–layer. We derive, analytically, the condition for marginal stability in an iono-sphere where the major variables vary exponentially, *albeit* with different scale lengths. We also discuss some aspects of the nonlinear evolution, and present results from the non-linear feedback onto the original perturbation. In our nonlinear analysis, we will remove a primary limitation in the work of *Chaturvedi and Ossakow* [1977], thereby obtaining a

self–consistent elementary nonlinear description of Rayleigh–Taylor modes.

2. Equilibrium and Linear theory

We begin by considering a plasma consisting of electrons and a single species of ions (typically O^+) in a constant magnetic field, $\mathbf{B} = \hat{e}_z B$ under the influence of gravity $\mathbf{g} = -g\hat{e}_x$ and a vertical electric field $\mathbf{E} = E(x)\hat{e}_x$. This field gives rise to an East–West ("zonal") motion, due to $\mathbf{E} \times \mathbf{B}$ drift. The discussion of a zonal electric field, related to the slow rise an fall of the ionosphere can be found in *Keskinen et al.* [1981]. We will operate in the equilibrium frame of the neutral wind, thus our "gravity" should be interpreted as an effective gravity, namely $g = g_0 - \Omega_i U_W$ where g_0 is the strength of the Earth's gravitational field, Ω_i the local ion cyclotron frequency, and U_W the eastward speed of the neutral wind (could be a negative quantity). Note that g is a positive quantity by convention, hence $\hat{e}_x$ is in the upward direction. We use the electrostatic approximation, appropriate for the low temperature ($T \sim 0.1\mathrm{eV}$) ionospheric plasma and for perturbations with wavelengths longer than the collisionless skin depth ($d_e = c/\omega_{pe} \sim 8\mathrm{m}$ for $n_e \sim 5 \ 10^5 \ \mathrm{cm}^{-3}$): $\mathbf{E} = -\nabla\phi$. We note in passing that this skin depth is of the same order as an ionospheric ion Larmor radius ($\rho_i = \sqrt{T_i/m_i}/\Omega_i \sim 4\mathrm{m}$ for $T \sim 0.1\mathrm{eV}$ and $B = 0.35$ Gauss), thus analyses dealing with wavelength in the 10m range and lower should include **both** thermal effects and coupling to the electromagnetic wave. The working equations are:

$$\frac{\partial n_j}{\partial t} + \nabla \cdot (n_j \mathbf{v}_j) = -\nu_R (n_j - n_{j0}) \tag{1}$$

$$0 \approx \mathbf{E} + \frac{1}{c}\mathbf{v}_e \times \mathbf{B} \tag{2}$$

$$m_i \left(\frac{\partial \mathbf{v}_i}{\partial t} + \mathbf{v}_i \cdot \nabla \mathbf{v}_i \right) = e\mathbf{E} + \frac{e}{c}\mathbf{v}_i \times \mathbf{B} - m_i \nu_{in}\mathbf{v}_i + m_i \mathbf{g} \tag{3}$$

$$\nabla \cdot \mathbf{E} = 4\pi e (n_i - n_e) \tag{4}$$

The first equation is that of continuity for each species, $j = e, i$ with the right hand side indicating the loss due to recombination (n_{j0} is the equilibrium density). Note that we neglect electron inertia , cf. Eq. (2), while recognizing that the choice of a vertical, inhomogeneous electric field makes the plasma slightly non–neutral, cf. Eq. (4). These equations do not include the effect of parallel dynamics; we will return to that later. Also note that the ion momentum conservation equation, Eq. (3), is missing the effect

of electron–ion collisions; this term would simply add a constant to the polarization drift, leaving linear stability unaffected.

Under the assumption that the ion cyclotron frequency is much larger than any other frequency in this problem, we solve for the electron and ion velocities:

$$\mathbf{v}_e = \frac{c}{B}\mathbf{E} \times \hat{e}_z \tag{5}$$

$$\mathbf{v}_i = \frac{c}{B}\left(\mathbf{E} + \frac{m_i}{e}\mathbf{g}\right) \times \hat{e}_z$$
$$+\frac{1}{\Omega_i}\frac{c}{B}\left[\frac{\partial}{\partial t} + \nu_{in} + \frac{c}{B}\left(\mathbf{E} + \frac{m_i}{e}\mathbf{g}\right) \times \hat{e}_z \cdot \nabla\right]\left(\mathbf{E} + \frac{m_i}{e}\mathbf{g}\right)_{\perp} \tag{6}$$

where the $1/\Omega_i$ correction to the $\mathbf{E} \times \mathbf{B}$ drift for the ions is known as the polarization drift and gives rise to the Pedersen current. Though the polarization drift is small compared to the $\mathbf{E} \times \mathbf{B}$ drift, the instability is driven by the combination of free energy (the density gradient) and charge separation (the polarization drift $\rightarrow$ Pedersen current). After some algebra, we obtain the dispersion equation:

$$\left(1 - \frac{v_A^2}{c^2}\frac{\omega - k_y v_0}{\omega + i\nu_{in} - k_y v_0}\right)\left(\frac{\partial^2}{\partial x^2} - k_y^2\right)\phi_1 + \left(\frac{i\nu_{in}'}{\omega + i\nu_{in} - k_y v_0} + \frac{n_i'}{n_i}\right)\frac{\partial \phi_1}{\partial x}$$
$$+ \frac{k_y}{\omega + i\nu_{in} - k_y v_0}\left(v_E'' + v_E'\frac{n_i'}{n_i} - \frac{k_y g}{\omega + i\nu_R - k_y v_E}\frac{n_i'}{n_i} - \frac{v_A^2}{c^2}\frac{\omega v_E''}{\omega + i\nu_R - k_y v_E}\right)\phi_1 = 0 \tag{7}$$

The prime and double–prime symbols refer to first and second derivatives of equilibrium quantities with respect to x (height), ϕ_1 refers to the perturbed electrostatic potential, $v_E = -cE/B$ is the equilibrium $\mathbf{E} \times \mathbf{B}$ drift, and $v_0 = v_E + g/\Omega_i \approx v_E$ is the y–component of the ion equilibrium drift. The two terms proportional to v_A^2/c^2, where $v_A = B/\sqrt{4\pi n_i m_i}$ is the Alfvén speed, are due to the non–neutrality of the plasma, itself a consequence of $\nabla \cdot \mathbf{E} \neq 0$ in the equilibrium. This coefficient is of order 2–3 10^{-3} in the ionosphere, thus generally negligible. This dispersion relation yields results in agreement with those of, e.g., *Satyanarayana et al.* [1984] and *Hudson and Kennel* [1975a]. From e.g. *Fejer* [1981] we estimate $v_E' \sim 5\ 10^{-4}$ sec-1 near 20:00 local time and conclude that it gives an unimportant correction to equatorial dynamics. We omit velocity shear effects hereafter and work in the reference frame of the plasma, $\omega \rightarrow \omega - k_y v_E$.

A. Exponential plasma density profile and parallel dynamics

Assuming an exponentially stratified plasma, $n'/n = 1/L = $ constant, and neglecting the spatial variation of ν_{in}, ν_R for simplicity, this equation admits the following solution:

$$\phi_1 \propto \frac{1}{\sqrt{n(x)}} \exp\left(ik_x x + ik_y y - i\omega t\right) \tag{8}$$

with eigenvalue:

$$\omega = -\frac{i}{2}\left(\nu_{in} + \nu_R\right)\left\{1 \pm \left[1 + 4\frac{\kappa^2 g/L - \nu_i \nu_R}{(\nu_{in} + \nu_R)^2}\right]^{1/2}\right\} \tag{9}$$

where $\kappa^2 \equiv k_y^2/\left(k_x^2 + k_y^2 + 1/4L^2\right)$. For $\nu_{in} > \left(\nu_R, \sqrt{g/L}\right)$ and $k_y^2 >> k_x^2 >> 1/L^2$, we recover the well-know result: $\gamma = g/L\nu_{in} - \nu_R$. Note that Eq. (8) is an **exact** solution of Eq.(7) for constant ν_{in}, ν_R and exponential density profile; it does not make any use of a local approximation. Also $n_1 \approx i\left(k_y c/\gamma B\right)\left(dn/dx\right)\phi_1 \propto \sqrt{n}$; this factor represents the expansion of the bubble as the disturbance moves up in the stratified atmosphere ($\gamma = \text{Im}(\omega)$ is the linear growth rate). In addition, this nonlocal solution has the property that $\gamma \to 0$ as $k_{x,y} \to 0$, because of the $1/4L^2$ factor. This removes an annoying feature of the local solution which has $\gamma \to$ constant (a feature common to all interchange–like modes, cf. *Chaturvedi and Ossakow* [1978]).

The inclusion of parallel dynamics is done by considering the parallel component of the momentum conservation equation for electrons:

$$\left(\frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla + \nu_{en} + \nu_{ei}\right)v_{ez} = -\frac{e}{m_e}E_z + \nu_{ei}v_{iz} \tag{10}$$

where now $E_z = -ik_\parallel \phi_1$. Since the ion parallel velocity is down by a factor of the mass ratio, $v_{iz} \approx \left(m_e/m_i\right)v_{ez}$, it can be neglected. One can then easily compute the contribution $(\partial/\partial z)nv_{ez}$ to the electron continuity equation and obtain the modified dispersion equation:

$$\left(\frac{\partial^2}{\partial x^2} - k_y^2\right)\phi_1 + \left(\frac{i\nu_{in}'}{\omega + i\nu_{in}} + \frac{n'}{n}\right)\frac{\partial \phi_1}{\partial x}$$

$$+ \frac{1}{\omega + i\nu_{in}}\left(-\frac{n'}{n}\frac{k_y^2 g}{\omega + i\nu_R} + \frac{k_\parallel^2 |\Omega_e \Omega_i|}{\omega + i\bar{\nu}_e}\right)\phi_1 = 0 \tag{11}$$

where $\bar{\nu}_e = \nu_{ei} + \nu_{en}$. In the limit $\bar{\nu}_e >> \omega$ this equation is solved as before, yielding ($\alpha_\parallel \equiv \left(k_\parallel/k_\perp\right)^2 L|\Omega_i \Omega_e|/g$):

$$\omega = -\frac{i}{2}\left(\nu_{in} + \nu_R + \alpha_\parallel \frac{\kappa^2 g}{L\bar{\nu}_e}\right)\left\{1 \pm \left[1 + 4\frac{\left(\kappa^2 g/L\right)\left(1 - \alpha_\parallel \nu_R/\bar{\nu}_e\right) - \nu_i \nu_R}{\left(\nu_{in} + \nu_R + \alpha_\parallel \kappa^2 g/L\bar{\nu}_e\right)^2}\right]^{1/2}\right\} \tag{12}$$

Parallel electron dynamics contribute twice in a stabilizing manner: they directly reduce the drive term $\kappa^2 g/L$ and add to the collisional stabilization term. This effect may be estimated from $\alpha_{\parallel} \nu_R/\bar{\nu}_e \sim 8 \ 10^5 \ \nu_R/\bar{\nu}_e$ for a disturbance with $\lambda_{\perp} \sim 5\text{km}$, $\lambda_{\parallel} \sim R_E$ in a plasma with a 10 km density scale length, cf. *Basu and Coppi* [1983]. We conclude by mentioning that no new instability arises from parallel electron dynamics (the solution given above was obtained in the limit of high electron collision frequency). This proof involves the use of Nyquist diagrams and is given in Appendix A.

B. Wavepackets in regions of local maximum density gradient

In this section, we adopt a description where the equilibrium plasma density is assumed to have a maximum gradient at some specific height, cf. *Perkins and Doles* [1975]. Thus the coordinate x is intended to represent the distance away from the reference height x_0. We then take $n'/n = (1/L)\left(1 - x^2/d^2\right)$ in Eq. (8), with $d << L$, and proceed with the analysis of the equation. Let $\phi_1(x) = n^{-1/2} \exp\left(-i\Theta x/2L\right) F(x)$ where $\Theta = \nu_{in}' L/\left(\omega + \nu_{in}\right) \approx$ constant (since it applies to the variation of the neutral density and the mode is localized over a distance comparable to d instead of L). The equation for $F(x)$ is:

$$\frac{d^2 F}{dx^2} - \left[\frac{1}{4}\left(\frac{\Theta}{L}\right)^2 + \frac{i}{2}\frac{\Theta}{L}\frac{n'}{n} + \frac{1}{2}\frac{n''}{n} - \frac{1}{4}\left(\frac{n'}{n}\right)^2 + k_y^2\left(1 + \frac{gn'/n}{\omega\left(\omega + i\nu_{in}\right)}\right)\right] F \qquad (13)$$

where we omit ν_R and $k_{\parallel}$ for simplicity; they can be added to the final result simply by comparison with Eq. (12). Now, for modes with wavelengths short compared to the principal scale, L, one need keep only the second and last terms from within the square brackets. The resultant equation can be solved in terms of Hermite functions:

$$F(x) = \exp\left(-\frac{\alpha^2 x^2}{2d^2}\right) H_l\left(\frac{\alpha x}{d}\right), \qquad \alpha^4 \equiv -k_y^2 d^2 \frac{g/L}{\omega\left(\omega + i\nu_{in}\right)} + \frac{i}{2}\frac{d^2\nu_{in}'/L}{\omega + \nu_{in}} \qquad (14)$$

and the eigenvalue condition: $0 = 2l + 1 + \left(k_y^2 d^2/\alpha^2\right) - \alpha^2$. The solution is:

$$\alpha^2 = \frac{1}{2}(2l + 1) + \sqrt{\frac{1}{4}(2l+1)^2 + k_y^2 d^2} \qquad (15-a)$$

$$\omega = -\frac{i}{2}\left(\nu_{in} - \frac{d^2\nu_{in}'}{2L\alpha^4}\right)\left\{1 \pm \left[1 + \frac{k_y^2 d^2}{\alpha^4}\frac{4g/L}{\left(\nu_{in} - d^2\nu_{in}'/2L\alpha^4\right)^2}\right]^{1/2}\right\} \qquad (15-b)$$

Since $\nu_{in}' < 0$ typically (the density of the neutral atmosphere decreases with height), there is an additional stabilization from this factor. Once again, the instability is proportional

to $\sqrt{k_y^2 / \left(k_y^2 + \langle k_x^2 \rangle \right)}$ where the factor $2l + 1$ plays the role of a dimensionless wavenumber in the x–direction. This result has been reported by *Basu and Coppi* [1984] (our expression reduces to theirs in the limit $\nu'_{in} = 0$,and $k_y d >> 1$). Thus, wavepackets have instability rates that follow the same "rules" as the convectively amplified plane wave solutions of the preceding section.

C. Marginal stability

Often, all one is interested is the conditions for onset of Spread–F, since the dynamics quickly become nonlinear. We can make some analytic progress by taking the limit of high collisionality (valid for the F–layer): $\nu_{in} >> \omega$ and assuming that the ionosphere is stratified exponentially. In particular: $n(x) = n_0 \exp\left(x/L_n\right)$, $\nu_{in} = \nu_{in0} \exp\left(-x/L_\nu\right)$, $\nu_R = \nu_{R0} \exp\left(-x/L_R\right)$. The label "0" refers to a reference bottomside boundary (e.g., the top of the E–layer). At marginal stability, the mode is at rest in the plasma reference frame ($\omega = 0$), the governing equation is:

$$\left(\frac{\partial^2}{\partial x^2} - k_y^2 \right) \phi_1 + \left(\frac{\nu'_{in}}{\nu_{in}} + \frac{n'}{n} \right) \frac{\partial \phi_1}{\partial x} + \frac{n'}{n} \frac{k_y^2 g}{\nu_{in} \nu_R} \phi_1 \tag{16}$$

This equation reduces to Bessel's equation, after several transformations, with solution:

$$\phi_1(x) = \hat{\phi}_1 \exp\left(-\frac{x}{2\mathcal{L}} \right) J_p \left[\Delta \exp\left(\frac{lx}{2L_R} \right) \right] \tag{17}$$

with $1/\mathcal{L} \equiv 1/L_n - 1/L_\nu$, $p \equiv 2k_\perp L_R/l$, $l \equiv L_R/L_\nu + 1$, $k_\perp^2 \equiv k_y^2 + 1/4\mathcal{L}^2$, $\Delta \equiv (2/l)\sqrt{k_y^2 L_R^2 g/L_n \nu_{in0} \nu_{R0}}$. It can be seen that this solution is well behaved, i.e. goes to zero as $x \to \infty \gg p$. This eigenfunction yields striations (cylindrical depletions) that are stacked non–uniformly: they pack closer together as one goes up in height.

The boundary condition at the bottom ($x = 0$) **on the nightside** is that of an insulator (the E–region has vanished): $J_x(x = 0) = 0$; this immediately yields $E_x(0) = \partial \phi_1 / \partial x|_0 = 0$:

$$\left(\frac{L_R}{2\mathcal{L}} + k_\perp L_R \right) J_p(\Delta) - \frac{l}{2} \Delta J_{p-1}(\Delta) = 0 \tag{18}$$

an equation that can be solved using standard tables. In the limit $p \gg 1$ we obtain: $k_y^2 g/L\nu_{in0}\nu_{R0} = \left(k_\perp + 1/4\mathcal{L} \right)^2$, reminiscent of the result from the local theory. In the opposite limit, $p \ll 1$, we obtain $\Delta \approx 2.5$–3.8 (for a wide range of $L_r/l\mathcal{L}$), which yields: $k_y^2 L_R^2 g/L_n \nu_{in0}\nu_{R0} \approx 6.0$–$14.5$ (in essence, impossible to achieve under ordinary ionospheric

conditions). These two estimates represent opposite (extreme) limits in the spectrum of perpendicular wavelengths and exemplify the selection of unstable modes by the combination of boundary conditions and differing scale lengths (of which there are three, L_n, L_ν, L_R). On the dayside the E–region is present, shorting out the perturbation, so that there is no boundary condition and no well–posed statement concerning marginal stability.

3. Nonlinear dynamics

As discussed in the introduction, there exist extensive numerical codes that can provide a description of the nonlinear dynamics. It is however always instructive to attempt to solve the problem analytically, since the answer is no longer dependent on case studies. In their seminal paper, *Rognlien and Weinstock* [1974] (and, later, *Chaturvedi and Ossakow* [1977]) recognized that a linearly unstable eigenmode will generate a stable "vertical" wave by direct coupling. In other words, the nonlinear harmonic has $k_y = 0$, and thus is predicted to be damped, cf. Eq. (9); it "costs energy" for the initial disturbance to generate this harmonic, leading to the eventual saturation of this instability. This has served as a useful description by most analytic nonlinear analyses of the Rayleigh–Taylor instability which have been carried out in the local approximation. However, as will be evident from what follows, a self–consistent calculation must take into account three other harmonics, in the general description of the nonlinearity. These nonlinear harmonics, four in total (not counting complex conjugates), arise from the self–interaction of a system of degenerate linear eigenmodes; the linear dispersion relation is a function of k_x^2, k_y^2, and k_z^2 (though we will see that parallel dynamics contribute negligibly to the nonlinearity). Since the modes are basically dispersionless, $Re(\omega) \approx k_y v_E$, the linear modes with $\pm k_x$ and $\pm k_y$ are equivalent and all possible interactions must be taken into account, cf. *Sudan and Keskinen* [1984].

A. The case of a plasma with an exponential density profile

We recall that, for the case of an exponential atmosphere (and roughly constant ion–neutral collision frequency), the problem admits an exact solution, cf. Eq. (8), namely that of a convectively modified plane wave. The eigensolution is that of a mode basically at rest in the plasma so that $\gamma_l \approx g/L\nu_{in} - \nu_R$ in the simplest limit (γ_l is the **linear** growth rate of the perturbation being analyzed) and $\phi_1 \approx -(iB/ck_y)\gamma_l L (n_1/n)$ $\propto [n(x)]^{-1/2} \exp(ik_x x + ik_y y + ik_\parallel z + \gamma_l t)$. Since the growth rate, cf. Eq. (9), is a function of k_x^2, k_y^2, k_z^2, there exist several combinations of wavevectors that yield the same stability properties. The coupling of all these different "basis" modes quickly becomes complicated. We have looked at the nonlinear forcing terms in the momentum equations, for the quasilinear and double–wavenumber harmonics (these correspond to the combinations $\mathbf{k} - \mathbf{k} = 0$ and $\mathbf{k} + \mathbf{k} = 2\mathbf{k}$ respectively) and found that the resultant terms are small compared to those contributed by the primary nonlinearity, due to the combination $(c/B)\mathbf{E} \times \hat{e}_z \cdot \nabla n$. One may also see this simply by realizing that the nonlinear velocity is $\mathbf{v}_2 = (1/\Omega_i)\,\hat{e}_z \times \mathbf{v}_1 \cdot \nabla \mathbf{v}_1$ (for the ions) and is down by an order γ_l/Ω_i compared to the primary nonlinearity. The electrons contribute terms of the same order ($\sim 10^{-4}$), typically of the form $\left(k_\parallel^4 L^2/k_\perp^2\right)\left(\gamma_l^2 \Omega_e^2/\bar{\nu}_e^3\right) n_1 |n_1/n|^2$. We thus keep only the lowest order $\mathbf{E} \times \mathbf{B}$ velocity in the nonlinear terms.

The symmetry properties of the linear solution indicates that there exist four "basis" modes, all equivalent: (k_x, k_y), $(k_x, -k_y)$, $(-k_x, -k_y)$, $(-k_x, k_y)$. However, the latter two are the complex conjugates of the first two. Hence only two basis waves (and their complex conjugates) need to be considered, representing specific choices of propagation. The intent here is to determine the nonlinear harmonics generated by the two–wave combination of 2 basis waves, then use another two–wave combination of the nonlinear harmonics combined with the basis waves to generate a forcing term for the original "linear" equation (collapsing this description to a single step yields the familiar three–wave coupling). The selection rules governing the phases generates only a finite number of combinations. Denoting by F (fundamental) the wave whose phase factor contains the combination (k_x, k_y), the following applies (where $\oplus$ denotes a wave coupling process):

$$(k_x, k_y) \oplus (-k_x, -k_y) \to (0, 0) \oplus (k_x, k_y) \to (k_x, k_y) \tag{19-a}$$

$$(k_x, k_y) \oplus (k_x, k_y) \to (2k_x, 2k_y) \oplus (-k_x, -k_y) \to (k_x, k_y) \tag{19-b}$$

$$(k_x, k_y) \oplus (-k_x, k_y) \to (0, 2k_y) \oplus (k_x, -k_y) \to (k_x, k_y) \tag{19-c}$$

$$(k_x, k_y) \oplus (k_x, -k_y) \to (2k_x, 0) \oplus (-k_x, k_y) \to (k_x, k_y) \qquad (19-d)$$

It is the last combination, one involving a zero k_y nonlinear mode, that was considered by previous authors, e.g. *Chaturvedi and Ossakow* [1977], *Hudson* [1978]. Other terms, such as the quasilinear, Eq. (19-a), or double–wavenumber, Eq. (19-b), were tacitly ignored. The other "basis" mode which we denote M (mixed) consists of the combination $(k_x, -k_y)$ and is subject to the following third order feedback:

$$(0, -2k_y) \oplus (k_x, k_y) \to (k_x, -k_y) \qquad (20-a)$$

$$(2k_x, 0) \oplus (-k_x, -k_y) \to (k_x, -k_y) \qquad (20-b)$$

The resulting equations for the evolution of the individual "basis" and nonlinear harmonics (subscript "2") are:

$$\frac{\partial n_{22}}{\partial t} = \Gamma_{22} n_{22} + \gamma_l \frac{n_F^2}{n} \qquad (21-a)$$

$$\frac{\partial n_{2V}}{\partial t} = \Gamma_{2V} n_{2V} + 4i\gamma_l k_x L \frac{n_F n_M}{n} \qquad (21-b)$$

$$\frac{\partial n_{2H}}{\partial t} = \Gamma_{2H} n_{2H} + 2\gamma_l \frac{n_F n_M^*}{n} \qquad (21-c)$$

$$\frac{\partial n_F}{\partial t} = \Gamma_F n_F - \gamma_l \frac{n_{22} n_F^*}{n} + \gamma_l L \left(\frac{n_M^*}{n} \frac{\partial n_{2V}}{\partial x} + \frac{n_M}{n} \frac{\partial n_{2H}}{\partial x} + n_{2H} \frac{\partial}{\partial x} \frac{n_M}{n} \right) \qquad (21-d)$$

$$\frac{\partial n_M}{\partial t} = \Gamma_M n_M + \gamma_l L \left(\frac{n_F^*}{n} \frac{\partial n_{2V}}{\partial x} + \frac{n_F}{n} \frac{\partial n_{2H}^*}{\partial x} + n_{2H}^* \frac{\partial}{\partial x} \frac{n_F}{n} \right) \qquad (21-e)$$

These equations are the equations for the three nonlinear harmonics $(2k_x, 2k_y)$, $(2k_x, 0)$, and $(0, 2k_y)$ (labelled "22", "2V", and "2H", respectively) and the third order equations for the two basis harmonics. The coefficients Γ_j are short–hand representations of the homogeneous terms in the continuity equation (involving operators), and correspond to the respective linear growth rates, so that $\Gamma_F = \Gamma_M = \gamma_l$ and, e.g., $\Gamma_{2V} = -\nu_R$. We should mention that the quasilinear harmonic, (0,0), is found to be time–independent, so that its amplitude can be taken equal to zero in this calculation. From our knowledge of the linear eigenfunctions, we quickly find that $\partial n_{2H}/\partial x = 0$ and $\partial n_{2V}/\partial x = 2ik_x n_{2V}$ by inspection of Eqs. (21-b), (21-c). One can then obtain a set of five ordinary differential equations for the dimensionless amplitudes, which in general are complex

$$\frac{\partial N_{22}}{\partial t} = \Gamma_{22} N_{22} + \gamma_l N_F^2 \qquad (22-a)$$

$$\frac{\partial N_{2V}}{\partial t} = \Gamma_{2V} N_{2V} + 4i\gamma_l k_x L N_F N_M \tag{22 - b}$$

$$\frac{\partial N_{2H}}{\partial t} = \Gamma_{2H} N_{2H} + 2\gamma_l N_F N_M^* \tag{22 - c}$$

$$\frac{\partial N_F}{\partial t} = \gamma_l N_F + \gamma_l \left[-N_{22} N_F^* + +2i k_x L N_{2V} N_M^* + \left(i k_x L - \frac{1}{2} \right) N_{2H} N_M \right] \tag{22 - d}$$

$$\frac{\partial N_M}{\partial t} = \gamma_l N_M + \gamma_l \left[2i k_x L N_{2V} N_F^* + \left(i k_x L - \frac{1}{2} \right) N_{2H}^* N_F \right] \tag{22 - e}$$

and we have set $\sqrt{n(x)/n(0)} \approx 1$ in the last two equations. The saturated state is easily obtained; the three nonlinear harmonics end up with the following amplitudes:

$$N_{22} = -\frac{\gamma_l}{\Gamma_{22}} N_F^2, \qquad N_{2V} = -4i\frac{\gamma_l}{\Gamma_{2V}} k_x L N_F N_M, \qquad N_{2H} = -2\frac{\gamma_l}{\Gamma_{2H}} N_F N_M^* \tag{23}$$

The two basis modes saturate with the following amplitudes:

$$|N_F|^2 = -\frac{\Gamma_{2V}}{\gamma_l} \left(8k_x^2 L^2 + \frac{\Gamma_{2V}}{\Gamma_{2H}} \right)^{-1} \tag{24 - a}$$

$$|N_M|^2 = -\frac{\Gamma_{2V}}{\gamma_l} \left(8k_x^2 L^2 + \frac{\Gamma_{2V}}{\Gamma_{2H}} - \frac{\Gamma_{2V}}{\Gamma_{22}} \right) \left(8k_x^2 L^2 + \frac{\Gamma_{2V}}{\Gamma_{2H}} \right)^{-2} \tag{24 - b}$$

and are subject to small, though finite, frequency shifts:

$$\omega_F = -2k_x L \frac{\gamma_l^2}{\Gamma_{2H}} |N_M|^2, \qquad \omega_M = -2k_x L \frac{\gamma_l^2}{\Gamma_{2H}} |N_F|^2 \tag{25}$$

Thus, the saturated state is completely specified through the third order description of a "basis" set of linearly unstable eigenmodes feeding back on itself. Note that, since Γ_{2V} is the only negative term while $\Gamma_F = \Gamma_M \approx \gamma_l > \Gamma_{2H} \geq \Gamma_{22} > 0 > \Gamma_{2V}$, we predict that the basis harmonics will saturate with unequal amplitudes: $|N_F|^2 \geq |N_M|^2$. Also saturation apparently requires finite wavenumber in the x–direction: $k_x^2 L^2 \geq -\Gamma_{2V}/8\Gamma_{2H} > 0$, a very mild condition. These convectively amplified plane waves will saturate with $|N_{2V}| \sim 1/2k_x L$ (when $k_x L > 1$), and $|N_F| \sim |N_M| \sim \sqrt{\nu_R/2\gamma_l} |N_{2V}|$ so that the ratio of amplitudes between the "linear" and vertical harmonics depends on the strength of the linear growth rate, i.e. is less than unity for the most unstable modes. These results are in agreement with *Chaturvedi and Ossakow* [1977].

B. Dynamics of wavepackets in regions of local density gradient maxima

We recall, from Sec. 1B, that the linear eigenmode is now a standing mode in the x–direction, cf. Eq. (14). Unlike the previous section, where two "basis" eigenmodes had to be taken into account (due to degeneracy in $\pm k_x$), here the situation is simpler: there exit only one fundamental eigenfunction, which generates "quasilinear" and "double–wavenumber" nonlinearities. The only other major difference with the previous section is that, here, there is a non–vanishing quasilinear density change. Working through some algebra, we obtain:

$$\frac{\partial \langle n_2 \rangle}{\partial t} = \Gamma_{20} \langle n_2 \rangle + 4\gamma_l \alpha \left(2l \frac{H_{l-1}}{H_l} - \frac{\alpha x}{d} \right) \frac{|n_1|^2}{n} \qquad (26-a)$$

$$\frac{\partial n_{22}}{\partial t} = \Gamma_{22} n_{22} + \gamma_l \frac{n_1^2}{n} \left(1 - 2\frac{xL}{d^2} \right) \qquad (26-b)$$

$$\frac{\partial n_1}{\partial t} = \Gamma_1 n_1 - \gamma_l \left[2\alpha \frac{L}{d} \left(\frac{x}{d} - \frac{l H_{l-1}}{H_l} \right) n_{22} \frac{n_1^*}{n} \right. \\ \left. - \frac{n_1^*}{n} \frac{\partial n_{22}}{\partial x} - \frac{n_1}{n} \frac{\partial \langle n_2 \rangle}{\partial x} \right] \qquad (26-c)$$

where we denote the quasilinear (zero k_y) density by $\langle n_2 \rangle$ and the double–wavenumber (nonlinear) density by n_{22}. Now, the terms Γ_j must be interpreted in terms of the original x–dependent operators and require us to solve a set of partial differential equations. We can make some qualitative progress, however, by considering the fundamental ($l = 0$) mode, replacing the Γ_j's by their linear values (γ_j, in particular $\Gamma_{20} = 0$) and realizing that then Eq. (26-a) predicts that the quasilinear term is damped in time from its original (infinitesimal) amplitude. Thus, one can set $\langle n_2 \rangle \approx 0$. We now estimate that saturation occurs when the wavepacket evolves by rising to a height $x \leq d$ above the region of maximum gradient. From Eq. (26-b) we obtain an estimate for the second nonlinearity: $n_{22} \approx (\gamma_l / \Gamma_{22}) (2L/d - 1) n_1^2/n$ (this requires that the bubble "lift", $x = d$ for saturation, since $\gamma_1 \geq \Gamma_{22} > 0$). The dominant terms in Eq. (26-c) are the first and second on the right side, thus:

$$\left| \frac{n_1}{n} \right|^2 \approx \frac{\Gamma_{22}}{\gamma_l} \frac{d/2L}{4\alpha - 1} \qquad (27)$$

indicating that these depletions will saturate with very small amplitudes since $\alpha > 1$, cf. Eq. (15-a), and $d/L < 1$. Physically, the bubble move up, away from the region where its growth is optimal, pushing against the denser fluid. This is energetically unfavorable and the process saturates.

4. Conclusions

We have analyzed the linear and nonlinear dynamics of collisional plasmas in the F–region of the equatorial ionosphere, concentrating on processes dominated by gravity and the density gradient of the plasma. By adopting two complementary descriptions of the plasma equilibrium, we have been able to give an analytic description of the problem, **without** making recourse to the local approximation. Thus, our results are relevant to wavelength regime, $\lambda_\perp \geq 1$ km where the spectrum of spread–F is maximum. Our findings are as follow:

i) Parallel electron dynamics contribute a strongly stabilizing term to the dispersion relation, even in the simplest case of electrostatic perturbations in a plasma with a homogeneous magnetic field. This stabilizing term scales with $L\lambda_\perp^2/\lambda_\parallel^2$ where L is the density gradient scale length. This is completely consistent with *Basu and Coppi* [1983], [1984] as well as *Glassman and Sperling* [1983]. The analysis presented is non–local and predicts a "deepening" of the depletion for upward propagating disturbances.

ii) A criterion for onset of Spread–F has been derived analytically, for an ionosphere with three different scale lengths (related to n, ν_{in}, and ν_R). It naturally provides a selection rule for the perpendicular wavelengths of unstable modes, cf. Eq. (18).

iii) For the simple case of a plasma with an exponential density profile we have generalized the three–wave coupling description of e.g., *Chaturvedi and Ossakow* [1977] and found expressions for the saturated amplitudes and nonlinear frequency shifts for the various harmonics. This has been done for the case of a single linearly unstable mode with given $|k_x|$, $|k_y|$. This description is correct if one wishes to examine the evolution of a large "longitudinal slice" of the F-region, so that coupling between different $|k_y|$ linear eigenmodes averages out. These disturbances will propagate up from the bottomside layer (the nonlinear model has constant ν_{in} and ν_R) as convectively amplifying plane waves, $n_1 \propto \sqrt{n(x)}$. There is, in principle, no "stopping height" for these disturbances, once they are launched.

iv) For the case of an ionosphere with one (or more) localized regions (pockets) of large density gradients, a solution consisting of standing wavepackets has been found. Their linear stability characteristics resemble those of the plane waves discussed above (as well as of the solutions commonly obtained under the "local approximation"), but their nonlinear dynamics are quite different: they propagate up only a finite distance (comparable to the shorter scale length, that describes the width of the original density

maximum) and saturate with much smaller amplitudes. Their eventual dynamics past that point are those of evanescent waves, although if these pockets are not too far apart from each other, it is possible to propagate the wavepackets over considerable distances.

Appendix A: Nyquist Diagram

The use of Nyquist diagrams in plasma physics dates from the well–know *Penrose* criterion [1960]. If we denote by $D(\omega) = 0$ the dispersion relation for the relevant perturbations, let ω trace a semi–circular contour in the upper–half of the complex plane. The integral of $\partial \ln D / \partial \omega$ is evaluated over this closed contour and the number of unstable roots (zeroes of D) is given by:

$$N = M + \frac{1}{2\pi i} \left[\int_\Gamma d\omega \, \frac{\partial \ln D}{\partial \omega} + \ln \frac{D(i\epsilon + \infty)}{D(i\epsilon - \infty)} \right]_{\epsilon \to 0} \qquad (A-1)$$

where N=number of zeros, $M=$ number of poles (of D), Γ is the arc (whose radius is taken to infinity) portion of the ω–contour, and the logarithm comes from integration along the straight segment (i.e., the real axis once $\epsilon \to 0$). The dispersion relation that arises from Eq. (11) reads:

$$D(\omega) = (\omega + i\nu_{in})(\omega + i\nu_R) + \frac{\kappa^2 g}{L} \left(1 - \alpha_\parallel \frac{\omega + i\nu_R}{\omega + i\bar{\nu}_e} \right) \qquad (A-2)$$

note that this equation admits no poles in the upper–half of the complex ω plane, thus M=0. The first part of the term within brackets in Eq. (A-1) is evaluated:

$$\int_\Gamma d\omega \, \frac{\partial \ln D}{\partial \omega} = \lim_{\rho \to \infty} i \int_0^\pi d\theta \, \rho \exp(i\theta) \frac{2}{\rho \exp(i\theta)} = 2\pi i \qquad (A-3)$$

since $D \approx \omega^2 + i\omega (\nu_{in} + \nu_R)$ on the arc Γ. There remains to be evaluated the second term in brackets. This logarithm is simply $2\pi i$ times the net number of up→down crossings of the negative real–$D(\omega)$ axis (i.e., the branch cut) as ω traces out the straight segment parallel to the real axis (actually, it is the real axis after we take the limit $\epsilon \to 0$). Hence, we do not need to obtain the entire $D(\omega)$ contour, but only the crossing points. From (A-2), we obtain that Im(D)=0 for

$$\omega^2 = \omega_R^2 \equiv \frac{\bar{\nu}_e \nu_R \nu_{in} - (\kappa^2 g/L)(\bar{\nu}_e + \alpha_\parallel \nu_R)}{\bar{\nu}_e - \nu_R - \nu_{in}} \qquad (A-4)$$

Note that no crossings are possible for either $\nu_R = 0$ or $\kappa^2 g/L > \nu_R \nu_{in}$ since ω is assumed real for this portion of the calculation. The corresponding real part of D is:

$$\mathrm{Re}(D) = \frac{\omega_R}{\left(\omega_R^2 + \bar{\nu}_e^2\right)\left(\bar{\nu}_e + \alpha_\parallel \nu_R\right)} \left\{ \omega_R^2 \left[\nu_R + \nu_{in} + \alpha_\parallel \left(\bar{\nu}_e - \nu_{in}\right)\right] + \bar{\nu}_e^2 \left(\nu_R + \nu_{in}\right) \right. $$
$$\left. + \alpha_\parallel \nu_R^2 \left(\bar{\nu}_e - \nu_{in}\right)\right\} \qquad (A-5)$$

Since $\bar{\nu}_e > \nu_{in}$, the term in brackets is positive definite and $sgn(D) = sgn(\omega_R)$. Hence, crossings of the real–D axis occur (when they do) in pairs on either side of the origin. Then the contours can be of only two kinds: (a) $\omega_R^2 < 0$, no crossings are possible, thus $N = 1 + 0$ and one unstable mode occurs, (b) $\omega_R^2 > 0$, there is one crossing of the branch cut coming from below, thus $N = 1 - 1$ and there are no unstable modes. This concludes our proof: there exists at most one unstable mode, the original Rayleigh–Taylor instability that is present when $k_\parallel = 0$.

Acknowledgments

References

Balsley, B. B. G. Haerendel, and R. A. Greenwald, Equatorial spread F: recent observations and a new interpretation, *J. Geophys. Res.* **77**, 5625, 1972.

Basu, B. and B. Coppi, Localized plasma depletion in the ionosphere and the equatorial spread F, *Geophys. Res. Lett.* **10**, 900, 1983.

Basu, B. and B. Coppi, Plasma depletion processes in the ionosphere and the equatorial spread F, MIT Report PTP-84/18 (1984).

Basu, Su., Sa. Basu, J. Aarons, J. P. McClure, and M. D. Cousins, On the co-existence of kilometer- and meter-scale irregularities in the nighttime equatorial F-region, *J. Geophys. Res.* **83**, 4219, 1978.

Chaturvedi, P. K. and S. L. Ossakow, Nonlinear theory of the collisional Rayleigh-Taylor instability in equatorial spread F, *Geophys. Res. Lett.* **4**, 558, 1977.

Chaturvedi, P. K. and S. L. Ossakow, High-altitude limit of the gradient drift instability, *J. Geophys. Res.* **83**, 2691 (1978).

Chaturvedi, P. K. and S. L. Ossakow, Nonlinear stabilization of the $\mathbf{E} \times \mathbf{B}$ gradient drift instability in ionospheric plasma clouds, *J. Geophys. Res.* **84**, 419, 1979a.

Chaturvedi, P. K. and S. L. Ossakow, Nonlinear stabilization of the current convective instability in the diffuse aurora, *Geophys. Res. Lett.* **12**, 957, 1979b.

Dungey, J. W., Convective diffusion in the equatorial F-region, *J. Atmos. Terr. Phys.* **9**, 304, 1956.

Fejer, B. G., The equatorial ionospheric electric field. A review, *J. Atmos. Terr. Phys* **43**, 377 (1981).

Fejer, B. G. and M. C. Kelley, Ionospheric irregularities, *Rev. Geophys. Space Phys.* **18**, 401, 1980.

Glassman, A. J. and J. L. Sperling, Electromagnetic theory of collisional interchange instabilities, *J. Geophys. Res.* **88**, 10091, 1983.

Guzdar, P. N., P. Satyanarayana, J. D. Huba, and S. L. Ossakow, Influence of velocity shear on the Rayleigh-Taylor instability, *Geophys. Res. Lett.* **9**, 547, 1982.

Haerendel, G., Theory of equatorial spread F, unpublished manuscript, Max-Planck Institut für Physik und Astrophysik, 1974.

Hassam, A. B., W. Hall, J. D. Huba, and M. J. Keskinen, Spectral characteristics of interchange turbulence, *J. Geophys. Res.* **91**, 13513, 1986.

Hoh, F. C., Instability of Penning-type discharges, *Phys. Fluids* **6**, 1184, 1963.

Huba, J. D., S. L. Ossakow, P. Satyanarayana, and P. N. Guzdar, Linear theory of the $\mathbf{E} \times \mathbf{B}$ instability with an inhomogeneous electric field, *J. Geophys. Res.* **88**, 425, 1983.

Hudson, M. K., Spread F bubbles: nonlinear Rayleigh–Taylor mode in two dimensions, *J. Geophys. Res.* **83**, 3189, 1978.

Hudson, M. K. and C. F. Kennel, Linear theory of Equatorial Spread F, *J. Geophys. Res.* **80**, 4581, 1975a.

Hudson, M. K. and C. F. Kennel, The electromagnetic interchange mode in a partly-ionized collisional plasma, *J. Plasma Phys.* **14**, 121, 1975b.

Kelley, M. C., The Earth's Ionosphere (Academic Press, New York, 1989), p. 114-115.

Keskinen, M. J., S. L. Ossakow, and P. K. Chaturvedi, Preliminary report of numerical simulations of intermediate wavelength collisional Rayleigh–Taylor instability in equatorial spread F, *J. Geophys. Res.* **85**, 1775, 1980.

Keskinen, M. J., E. P. Szuszczewicz, S. L. Ossakow, and J. C. Holmes, Nonlinear theory and experimental observations of the local collisional Rayleigh–Taylor instability in a descending equatorial spread F ionosphere, *J. Geophys. Res.* **86**, 5785, 1981.

Ossakow, S. L., Ionospheric irregularities, *Rev. Geophys. Space Phys.* **17**, 521, 1979.

Ossakow, S. L., Spread–F theories – A review, *J. Atmos. Terr. Phys.* **43**, 437, 1981.

Ossakow, S. L., S. T. Zalesak, B. E. McDonald, and P. K. Chaturvedi, Nonlinear equatorial spread F: dependence on the altitude of the F peak and bottomside background electron density gradient scale length, *J. Geophys. Res.* **84**, 17, 1979.

Ott, E., Theory of Rayleigh–Taylor bubbles in the equatorial ionosphere, *J. Geophys. Res.* **83**, 2066, 1978.

Penrose, O., Electrostatic instabilities of a uniform non–Maxwellian plasma, *Phys. Fluids* **3**, 258, 1960.

Perkins, F. W. and J. H. Doles, III, Velocity shear and the $\mathbf{E} \times \mathbf{B}$ instability, *J. Geophys. Res.* **80**, 211, 1975.

Perkins, F. W., N. J. Zabusky, and J. H. Doles, III, Deformation and Striation of Plasma Clouds in the Ionosphere, 1, *J. Geophys. Res.* **78**, 697, 1973.

Rognlien, T. D. and J. Weinstock, Theory of the nonlinear spectrum of the gradient drift instability in the equatorial electrojet, *J. Geophys. Res.* **79**, 4733, 1974.

Satyanarayana, P., P. N. Guzdar, J. D. Huba, and S. L. Ossakow, Rayleigh–Taylor instability in the presence of a stratified shear layer, *J. Geophys. Res.* **89**, 2945, 1984.

Simon, A., Instability of a partially ionized plasma in crossed electric and magnetic fields, *Phys. Fluids* **6**, 382, 1963.

Sudan, R. N. and M. J. Keskinen, Unified theory of the power spectrum of intermediate wavelength ionospheric electron density fluctuations, *J. Geophys. Res.* **89**, 9840, 1984.

PREPARATION AND CHARACTERIZATION
OF BLENDS OF POLY(METHYL METHACRYLATE)
WITH NOVEL SILSESQUIOXANE MATERIALS

Charles J. Noel
Associate Professor
Department of Consumer and Textile Sciences

The Ohio State University
1787 Neil Avenue
Columbus, OH 43210-1295

PREPARATION AND CHARACTERIZATION
OF BLENDS OF POLY(METHYL METHACRYLATE)
WITH NOVEL SILSESQUIOXANE MATERIALS

Charles J. Noel
Associate Professor
Department of Consumer and Textile Sciences
The Ohio State University

ABSTRACT

Polyhedral oligomeric silsesquioxane (POSS) materials are an important class of compounds having an oxygen/silicon ratio of 1.5, a value intermediate between that found in silicones and silica. An entirely new class of polymerizable monomers containing POSS entities is currently under development at Phillips Laboratory. Incorporation of POSS entities into conventional organic polymers shows promise of enhancing certain properties of these polymers, particularly thermal properties, and may extend the useful temperature range of these materials, and alter their flammability characteristics. In this study, the extensively used thermoplastic poly(methyl methacrylate) (PMMA) was blended with two different POSS compounds, one a POSS macromer, $3\text{-}T_8Cy_7\text{propyl}$ methacrylate, and the other a copolymer of methyl methacrylate with the same POSS macromer. Blends were prepared by dissolving the materials to be blended in a common solvent, thoroughly mixing the solution, and recovering the solid blend by evaporating the solvent. Procedures were developed for removing virtually all solvent from the blends, and for processing the blended materials into useable forms for further extensive characterization. Blends were characterized using conventional thermal analytical techniques, including Differential Scanning Calorimetry (DSC), Thermogravimetric Analysis (TGA), and Thermomechanical Analysis (TMA). Incorporating the POSS entity into a copolymer with methyl methacrylate greatly increases the amount of POSS which can be blended with PMMA. Clear films were obtained at the 50/50 w/w blend level; these blends contain 35% POSS by weight. When the homopolymer of the same POSS macromer was blended with PMMA, clear films could only be obtained with a POSS loading of 9% or less.

PREPARATION AND CHARACTERIZATION
OF BLENDS OF POLY(METHYL METHACRYLATE)
WITH NOVEL SILSESQUIOXANE MATERIALS

Charles J. Noel

INTRODUCTION

Blending of two different polymers with the aim of developing materials with enhanced properties continues to be important in the materials industry. Most of the blends reported in the literature are blends of two organic polymers, but some work on blends of an organic with an inorganic polymer have been reported. Allcock and co-workers (1,2) prepared blends of polyphosphazines with a variety of organic polymers. Their results showed that blends would retain miscibility over time if strong interactions were possible between the constituent polymers, but that phase separation would occur if only weak interactions were possible. Landry *et al.* report on novel blends of etheric polyphosphazines with organic polymers, indicating that miscibility is enhanced in these systems if the organic polymers contain acidic functionality. Other work in the literature is concerned with factors influencing the nature of the blends produced, Venugopal, Krause, and Wnek (4) report that electric fields applied during the solvent casting of polymeric blends from single-phase solutions affected the morphologies of the blends. Lu, Pearce, and Kwei (5) report that the miscibility window of blended polymers was affected by the choice of solvent used   They studied blends of a 50/50 w/w poly(styrene-co-4-vinylphenyl-dimethylsilanol) with poly(n-butyl methacrylate). The miscibility window was found to be 4 to 18 mole% of the copolymer when toluene was used as the casting solvent, compared to 9 to 34 mole% copolymer when 2-butanone was used. This work suggests that an equilibrium state of mixing is not always achieved in solvent casting films.

The work described in this report is a continuation of work carried out by the present author as an AFOSR Summer Faculty Fellow in the summer of 1994.   In that study (6) blends of poly(methyl methacrylate) and of a commercial thermoplastic elastomer (PEBAX 2533) with three polyhedral oligomeric silsesquioxane (POSS) materials were prepared and characterized using thermoanalytical techniques. Two of the POSS materials were inert monomeric compounds, fully cyclohexylated T_6 and T_8, while the third was poly(3-T_8Cy_7 propyl methacrylate). The T_8 and T_6 designations are shorthand for Si_8O_9 and Si_8O_{12}, respectively, while Cy stands for cyclohexyl, C_6H_{11}. POSS materials are an important class of compounds having an oxygen/silicon ratio of 1.5, a value intermediate between that found in silicones and silica. An entirely new class of polymerizable monomers containing POSS entities is currently under development at Phillips Laboratory; the chemistry of these materials has been described by Lichtenhan (7). In the 1994 work, blends of PMMA with poly(3-T_8Cy_7PMA) were prepared at levels of up to 9% (w/w) of the POSS polymer; phase separation was observed at higher levels of POSS polymer. Blending PMMA with the two inert POSS materials tended to produce films with the POSS concentrated

at the film surface, due to the greater solubility of the POSS materials in the solvent employed. The present study is concerned with the preparation and characterization of blends of PMMA with two POSS materials: monomeric 3-T_8Cy_7PMA and a copolymer of MMA and 3-T_8Cy_7PMA. The rationale for the use of these two materials is that inclusion of the methyl methacrylate functionality in the POSS materials might enhance their compatibility with PMMA by allowing stronger interactions between the blend components.

METHODOLOGY

1,   Blend Preparation

Blends were prepared by dissolving weighed amounts of materials in an appropriate solvent (chloroform), mixing thoroughly, and then allowing the solvent to evaporate slowly. A total weight of either 1.0 or 2.0g of solid material was dissolved in 20mL chloroform in 25mL scintillation vials. Solution was accomplished by a combination of techniques which included shaking the vials for several hours at room temperature using a wrist-action shaker, exposing the solutions to ultrasound at 45°C for up to two hours, and warming at 60°C for one to two hours. In all cases, clear single-phase solutions were obtained prior to solvent casting, The clear solutions were poured into aluminum weighing pans, and a 400mL beaker was inverted over each pan to allow evaporation to proceed slowly. All solvent casting was done in a fume hood. Solid materials were obtained after overnight evaporation in most cases. After casting, the solid films retained chloroform.

2.   Solvent Removal

Residual solvent has a significant effect upon many properties of the cast films. In some analyses, the solvent present can be compensated for, but it is far better that the solvent be removed. When a vacuum oven operated with house vacuum (18 to 20"Hg) at 105°C is used, some solvent is removed, and more is removed if the same oven is used with a mechanical vacuum pump (27"Hg). However, the amount of solvent removed was found to vary with the position of the sample in the oven, and removal of all solvent was never achieved. Two methods were found to achieve virtually complete solvent removal. Treatment of cast films with supercritical CO_2 in a pressure vessel modified so that it was swept continuously with the dense gas removed nearly all the chloroform in a thirty-minute exposure. Complete removal of chloroform from the cast films was achieved by processing the material in a high temperature press using the die as a modified vacuum oven. Two twenty-minute cycles were needed to achieve complete solvent removal. In the first cycle, vacuum was applied in the absence of pressure and then a disk was molded. In the second cycle, the material from the first disk was reused; vacuum was applied with pressure and a second disk was compression molded.

3.   Thermal Analysis

Characterization of materials was accomplished using three varieties of thermal analysis. *Thermomechanical Analysis (TMA)* provides a means of determining the softening temperature of the material being studied. A T.A Instruments Model 929 Thermo Mechanical Analyzer equipped with a Model 2000 Thermal Analyst Controller was used for this work. Materials studied were in powder form; these were compressed into a pellet approximately 1mm thick. The penetration probe with a 10g weight was centered on the pellet and the temperature was raised from room temperature to some predetermined elevated temperature at the rate of $3^{\circ}C/m$ in a flowing nitrogen atmosphere. The temperature at which the probe begins to penetrate the material is taken as the softening temperature.

Thermogravimetric Analysis (TGA) provides a means of following sample weight as its temperature is increased at a constant rate. A T.A. Instruments Model 951 equipped with a Model 2000 Controller was used. Samples weighing up to 20mg were placed on a platinum boat suspended on a quartz arm of a microbalance, and the temperature was raised from room temperature to $1000^{\circ}C$ at a nominal rate of $10^{\circ}C/m$ under an atmosphere of nitrogen gas flowing at 50mL/m. Both integral and time derivative weight-temperature traces were plotted and analyzed.

Differential Scanning Calorimetry (DSC) provides a means of observing first-order thermal transitions such as melting and crystallization as well as second-order transitions such as the glass transition temperature (T_g). A T.A. Instruments Model 912 equipped with a Model 2000 Controller and either a dual cell or a single cell with a liquid nitrogen cooling accessory was used. A sample weighing 5 to 15 mg was placed in a standard aluminum pan (weighing approximately 12 to 13mg) and an aluminum lid (9 to 10mg) was crimped onto the pan to cover the sample. An empty pan/lid combination served as the thermal reference; hermetically sealed pans were not used. DSC work reported here was carried out with material from which solvent had not been completely removed, so the samples were first heated from room temperature to $180^{\circ}C$ at the rate of $10^{\circ}C/m$ to expel residual solvent, cooled, then reheated to $180^{\circ}C$ at the rate of $10^{\circ}C/m$. If the single cell instrument equipped with the liquid nitrogen cooling accessory was used, the sample was cooled to $-100^{\circ}C$ at the rate of $-10^{\circ}C/m$. If the dual cell instrument was used, the sample was cooled manually with liquid nitrogen. The single cell instrument was observed to provide a better baseline than the dual cell instrument. All DSC runs were made in an atmosphere of nitrogen gas flowing at 50mL/m. Curves of heat flow through the samples relative to the reference as a function of temperature were recorded and analyzed. Data were recorded for both heating cycles in all cases, and were also obtained for the cooling cycle when the single cell instrument was used.

RESULTS

1.   <u>Characterization of Candidate Copolymers</u>

A series of copolymers of MMA with 3-T8Cy7PMA had been prepared (A. G. Bolf) for possible blending with PMMA. These ranged from 2.5 to 50 mole% (nominal) of the POSS comonomer. Homopolymers of both momomers were also prepared. These materials were characterized by both TMA and TGA. It should be noted that the *weight percent* of the POSS monomer in these copolymers ranges from 25.1% for the 2.5 mole% copolymer to 92.9% for the 50 mole% copolymer (assuming that the final copolymer composition is the same as the monomers used). The homopolymers and copolymers were provided in the form of white, free-flowing powders. For *TMA* work, quantities of material were pressed in a pellet press to form cylindrical pellets approximately 1mm high. It was noted that during the first heating of these pressed pellets, the pellets expanded thermally at an extremely high rate prior to the onset of probe penetration. Coefficients of Linear Thermal Expansion (CLTE) as high as 580 microns/meter °C were observed during the first heating of pressed pellets. This expansion was not reversible; the pellet height did not return to its original value on cooling. Subsequent heatings of the same pellets gave CLTE values of 50 to 70 microns/meter °C. The TMA trace for an extruded pellet of PMMA (as received from Rohm and Haas) showed a low CLTE and a softening temperature of 114°C, consistent with the literature. Pressed pellets from PMMA made by A. G. Bolf showed high values of CLTE for the first heat, lower values of CLTE on subsequent heatings, and softening temperatures of 136°C. The reason for the higher softening temperature is not known at this time; it may be related to the use of pressed pellets rather than molded or extruded material. Pressed pellets of the POSS homopolymer behaved similarly to the pressed PMMA pellet in that the CLTEs seen on first heats were much higher than those observed on second and third heats. However, a major difference was noted: the penetration probe never penetrated very deeply into the pellet when the preplanned temperature was reached. Further, when the preplanned temperature was shifted to higher values with each heating of a given pellet, penetration was observed to begin near the highest temperature previously experienced by the material. For example, if a given pressed pellet were heated three times, first to 200°C, second to 250°C, and third to 300°C, then penetration would be observed to begin below 200°C on the first heat, around 210°C on the second heat, and around 255°C on the third heat. Following the first heat, the material had the appearance of a sintered or fritted material, and this appearance did not change on subsequent heating. The pressed disks made from poly(3-T_8Cy_7PMA) never showed any evidence of their having flowed. It is doubtful that there is a true softening temperature for this material. One further note concerning this polymer: A small (8mg) fragment of a pressed disk after a second heating was placed in tetrahydrofuran. Material was observed peeling off the surface of the fragment. The remainder of the fragment dissolved in the THF, but the material which peeled off the surface swelled

and had not dissolved after ten weeks. This suggests surface cross-linking; since the pellets were both heated and cooled under flowing nitrogen, oxidation can probably be ruled out. Three of the A. G. Bolf copolymers were pressed into pellets and run. These were the 30, 20, and 10 nominal mole% POSS copolymers. The behavior of these three copolymers was similar to the behavior of the POSS homopolymer in that penetration was never complete, and the temperature for the onset of penetration seemed to depend upon thermal history, The copolymers also acquired the appearance of sintered materials, and did not appear to flow. Based upon these observations, the use of TMA with pressed pellets was discontinued, and no further TMA work is planned until compression molded disks can be used.

The candidate copolymers were also characterized by *TGA*. All were run, including both homopolymers, Since the materials were supplied as powders, they were used as received. No solvent was seen in any of the samples. The shape of the weight-temperature curve changed gradually in appearance as the nominal mole% of the POSS comonomer increased. PMMA is known to depolymerize almost quantitatively to monomer in a single step beginning near 280°C and ending around 400°C. Poly(3-T_8Cy_7PMA) decomposes at a considerably higher temperature than PMMA , with the first step beginning near 390°C, a second step beginning near 480°C, and a residue of approximately 37% at 1000°C. As the nominal mole% of the POSS comonomer increased, the temperature for 10% weight loss increased, the amount of weight lost in the higher temperature step increased, and the amount of residue at both 425°C and 1000°C increased. Plots of % residue and % weight lost in the high temperature step *vs.* mole% of the POSS comonomer showed rapid increases at low values of POSS mole% and gradual leveling at higher values of POSS mole%. These curves could all be fit to exponential functions of the form:

$$Y = A_0(1 - e^{-A_1 X})$$

where Y represents any of the weights, X is the mole% POSS comonomer, and A_0 and A_1 are parameters used to fit the data. The values of A_0 are close in numerical value to the corresponding experimental values for the POSS homopolymer. Higher values of A_1 show a more rapid approach to the asymptote, while lower values show a more gradual approach. Curve fitting was done by Dr. K. Chaffee, using the program GEN-PLOT. Chi-square tests for goodness of fit showed high levels of significance for all plots analyzed. The TGA data indicate that copolymerizing 3-(T_8Cy_7PMA) with MMA produces material with higher thermal stability than PMMA alone. At least 10 nominal mole% of the POSS comonomer should be used, and little is gained by exceeding 30 nominal mole%. Based on these thermal data, it was recommended that a nominal 70/30 mole% MMA/3-(T_8Cy_7PMA) copolymer be prepared at a 20g scale to provide an adequate supply of material for blending with PMMA.

2.   Blends of PMMA with POSS Macromer

A series of blends of PMMA with the unpolymerized 3-(T_8Cy_7PMA), occasionally referred to as a POSS *macromer* because it is a monomer with an exceedingly high molecular weight (1125.9), was prepared to see if enhanced thermal stability could be provided simply by using the macromer as a filler. Blends were prepared ranging from 100/0 to 70/30 w/w PMMA/POSS, in increments of 3% POSS material, by dissolving the appropriate weights of the two materials in chloroform then casting films in aluminum weighing pans. The resultant films were all clear, and were dried for two hours in a vacuum oven at 108°C and 17 to 18"Hg. After this treatment, the films remained clear, but had "popped" in many areas. These materials were characterized by TGA. All TGA traces showed an initial weight loss associated with thermal release of residual solvent, showing that the drying treatment was insufficient to remove all solvent. Solvent retained ranged from 5.37 to 16.73% by weight, and averaged 13.27%. There was no relationship between solvent retained and material composition. Following the solvent removal, weight loss appeared to occur in a single step for up to 12% w/w POSS material, with a small second step being observed at higher POSS material loadings. The first degradation step was characteristic of PMMA with respect to temperature, while the second degradation step occurred near 500°C , slightly below the temperature observed for the various poly(MMA-co-3-T_8Cy_7PMA) materials described above. When the %Weight Loss and %Residue at 1000°C observed were corrected for the solvent content, the %Weight Loss in the first degradation step decreased linearly with decreasing PMMA content, according to:

$$Y = 99.19 - 0.3534X \quad (r = 0.9248)$$

where Y = %Weight Loss in the first step, corrected for solvent, and X = weight fraction PMMA in the blend. For the 76/24 w/w blend, the sum of the corrected %Residue and the corrected %Weight unaccounted for (combining the corrected weights in the first degradation step and the residue) was 7.97%. This was quite close to the sum of the %Residue at 1000°C and %Weight Loss in the high temperature step observed for the nominal 97.5/2.5 m/m copolymer, 8.04%. This copolymer has a calculated composition of 74.9/25.1 w/w, assuming that the copolymer composition is the same as the charged monomers. At low levels of POSS macromer loading (up to 12% w/w) there is little enhancement of the thermal stability of PMMA; some enhancement is seen at higher loadings.

3.   Initial Blending of PMMA with 70/30 m/m Copolymer

A nominal 70/30 m/m copolymer of MMA with 3-T_8Cy_7PMA was synthesized at the 20g scale by A. G. Bolf. After purification and vacuum drying, a weight of 13.52g was recorded, corresponding to a yield of close to 68%. This material, designated AB-IV-28, was used in all subsequent blends with PMMA covered in this report. If the copolymer actually has the same composition as the charged

monomers, it should contain 82.8% of the POSS macromer by weight. Blends were prepared as above, by solvent casting from single-phase solutions in chloroform. Blend compositions ranged from 100/0 to 70/30 w/w PMMA/AB-IV-28, in increments of 3% POSS material; all blends were clear as cast. These blends were dried over the weekend in a vacuum oven at 108°C using a mechanical vacuum pump producing a vacuum of 27"Hg. These cast films had all extensively "popped" with the exception of the three nearest the front of the vacuum oven, which had only partially "popped", leaving large clear areas. These samples were all characterized by TGA, and all showed some residual solvent. Those which had extensively "popped" contained an average of 2.65% solvent, ranging from 0.58 to 5.70%; the three samples at the front of the oven contained an average of 11.72% solvent (10.21 to 12.52%). The TGA traces of the 100/0, 97/3, 96/4, and 91/9 blends showed no evidence of the high temperature degradation step. This step was seen, however, for all blends with higher copolymer content. In most cases, the weight loss associated with this step was small and most of the weight loss could be accounted for by the corrected lower temperature degradation step and the corrected 1000°C residue. Since the residue was lower than expected, a number of repeat runs were made with the higher POSS level blends. In some of these, larger weights associated with the high temperature step and the 1000°C residue were found, but in general, less POSS material than expected was found. These results suggest that either the two components of the copolymer are not uniformly distributed in the blends and that sampling the extensively "popped" material is a problem, or that the amount of the POSS comonomer in the copolymer is less than expected. Samples of AB-IV-28 were examined by proton NMR, and the ratio of the peak areas associated with the methyl in the ester group of MMA (3 Hydrogens) and the first methylene in the propyl ester group of $3\text{-}T_8Cy_7PMA$ (2 Hydrogens) was determined. If the copolymer has a 70/30 m/m MMA/$3\text{-}T_8Cy_7PMA$ composition, this ratio should be $(3 \times 7)/(2 \times 3) = 3.5$. Two runs with the AB-IV-28 copolymer gave ratio values of 7.8 and 6.0, corresponding to 16 and 20 mole% POSS in the copolymer. Based on these data, it appears that the composition of the copolymer AB-IV-28 is $82 \pm 2/18$ ± 2 m/m MMA/$3\text{-}T_8Cy_7PMA$. On this account, preparation of PMMA/copolymer blends for more extensive characterization was done extending the composition range of the blends to higher levels of the POSS material.

4.   Higher Level Blends of PMMA with AB-IV-28

Blends of PMMA with the copolymer AB-IV-28 were prepared at the 100/0, 80/20, 70/30, 60/40, and 50/50 w/w PMMA/AB-IV-28 blend levels. Four samples of each blend were prepared as described above, with 2.0g of total solids being used for each sample except for the 80/20 blend where four 1.0g samples were prepared. Following solvent casting from clear, single-phase solutions, the films were removed from their aluminum weighing pans and were weighed. All castings were clear. The four castings of each blend were broken into small pieces and were thoroughly combined to form five large

samples for further work. All samples were found to have retained considerable solvent:

Composition	Dry Weight	Cast Weight	%Solvent
100/0	8.055g	11.914g	47.9
80/20	4.063	5.178	27.4
70/30	8.098	11.289	39.4
60/40	8.153	11.273	38.3
50/50	8.063	10.953	35.8

In the data above, the %Solvent retained is expressed as a percentage of the dry weight of the blended material. Solvent retention is greater for the thicker 2.0g castings than for the thinner 1.0g castings. Solvent retention appears to decrease with decreasing PMMA content. Based upon the amount of solvent used to prepare the solutions for casting, about 90% of solvent used was removed by evaporation from the 2.0g castings, and about 96% from the 1.0g castings. Taking the copolymer composition to be 82/18 m/m $MMA/3-T_8Cy_7PMA$, the five materials contain 0.0, 14, 21, 28 and 35% of the POSS material by weight.

Samples taken from the 100/0 blend (all PMMA) were exposed to supercritical CO_2 by M. Carr, to see if the solvent could be removed in that manner. In the first experiment tried, a sample was exposed in a static reactor charged with CO_2 to 1050psi at 40°C and was left overnight. Under these conditions, above its critical point, CO_2 is a supercritical fluid, and might be expected to remove the solvent from the polymer blend. In the morning, the pressure was found to have fallen below critical during the night, due to a leak in the apparatus. When the sample was removed from the reactor, it was found to have changed from clear to a uniform white appearance, due to microbubbles having formed in the sample. TGAs run on both original and treated material showed 28.85% solvent in the untreated control, and 10.05% solvent in the treated material, based upon the total weight of the sample (including solvent). The leak in the apparatus was found and corrected, and a second sample was treated overnight at 1100psi and 65°C. This sample consisted of three clear strips of material. The pressure was found to have held overnight. When the sample was removed the three formerly clear strips of material were fused into a single white mass. TGA showed this material to contain 1.71% solvent, based upon total sample weight. A sample of the 60/40 blend was run with the reactor modified so that the sample could be swept continuously with $SCCO_2$ supplied from a heated bomb. This sample was extracted at 1250psi and 65°C for one hour. TGA showed that this sample contained 3.86% solvent, based on total sample weight. Clearly, extraction with $SCCO_2$ shows promise as a technique for solvent extraction from solvent-cast polymer films.

Work was also done, with P. Ruth, to attempt to remove solvent during the process of compression molding disks to be used for more extensive characterization studies. In the conventional compression molding process, material is placed in a die, a heavy plunger is placed over the material, the assembly is placed in a heated press, vacuum is applied to the die, the assembly is brought to temperature, the material melts and flows to the shape of the mold under the applied pressure. A solvent cast film of an older sample of PMMA was broken into small pieces and pressed into a 25mm disk in this manner. TGAs run on both the untreated control and the molded disk showed 17.99% solvent for the control, and 17.24% solvent for the molded disk. Solvent is not removed even though the PMMA flowed to form a clear disk. Experiments were carried out in which the variables of time, temperature and pressure were manipulated to determine how they influenced solvent removal. The best procedure devised was as follows:

1. Material is broken into small pieces and placed in the die.
2. The die is covered with a heavy temperature resistant rubber diaphragm.
3. The press is closed and the temperature raised to a preselected value.
4. Vacuum is applied to the die.
5. The material is held for 20 minutes at temperature and vacuum.
6. The press is opened and the diaphragm removed.
7. The pre-heated plunger is inserted in the die.
8. The press is closed and returned to temperature.
9. Pressure is applied to the material for 3.5 minutes, with vacuum.

This procedure was used to compression mold a disk from the 80/20 blend. TGAs on the untreated control and the molded disk showed 20.40% solvent for the control, and 2.50% solvent for the disk. The molded disk was broken into pieces and the procedure was repeated except that for the second treatment, the plunger was inserted in Step 2 instead of the diaphragm, and Steps 6, 7, and 8 are eliminated. TGA showed that this twice-molded disk had 0.00% solvent with no weight loss observed below 294°C. Disks were prepared from the 100/0 blend (all PMMA) at a temperature of 150°C, while higher temperatures (up to 220°C) are needed for blends containing the copolymer. Clear disks were prepared from the 100/0 blend (all PMMA). Clear disks were also prepared from each of the blends containing the AB-IV-28 copolymer, but these disks developed small internal cracks on cooling. Removing the disks from the mold while hot seems to reduce this cracking, as does cooling in the die under pressure.

5.   Characterization of Materials by DSC

When two polymers are blended, the usual criterion for determining that a single-phase blend is obtained is established by measuring T_g using DSC. If one phase is present, a single T_g is observed, at a temperature between the T_gs of the two polymers, and depending upon the composition of the blend. If the two polymers are phase separated, they behave independently of each other and two T_gs are observed. When poly(3-T_8Cy_7PMA) is used as one of the components of a blend, a difficulty arises because this material does not exhibit a T_g by DSC. DSC experiments run on the AB-IV-28 copolymer also showed no T_g. The absence of a Tg makes it extremely difficult to determine whether or not the blends are single-phase or phase separated. A variety of PMMA samples were examined by DSC, using a heat - cool - heat cycle. No consistent results were seen on the first heat to 180°C, but fairly consistent values of T_g were obtained on the second heat. Values found were between 114°C and 122°C, in fairly good agreement with published literature. Presumably the material flows during the first heat and makes better contact with the aluminum sample pan so the transition is clearly observed during the second heat. Two different nominal 70/30 copolymers, AB-IV-10 and AB-IV-28, were examined by DSC in the same manner. The earlier sample showed a T_g of 138°C on the second heat using both powder and a pellet pressed for TMA. The later sample, prepared specifically for the blend study, gave no indication of any T_g on second heats as high as 200°C. However, the same material, AB-IV-28, was able to be compression molded into very thin 12mm disks. The disks were clear but very brittle, and cracked on cooling. Several DSC runs were made on the first 70/30 w/w PMMA/AB-IV-28 blend. First heats gave horribly erratic traces, but in some cases, second heats gave clear indications of the existence of a Tg. Values were found between 117°C and 126°C, similar (or slightly higher?)to those values found for PMMA alone. In other runs with the same material, no Tg was observed for second heats. A more careful study using a DSC with a more stable baseline is warranted.

CONCLUSIONS

This study focused upon the thermal characterization of blends of PMMA with POSS entities containing methacrylate functionalities. Clear solvent-cast films and clear solvent-free compression molded disks were obtained from blends containing as high as 50% by weight of a copolymer of methyl methacrylate with 3-T_8Cy_7-propyl methacrylate. These blends contain up to 35% of the POSS material by weight. In an earlier study (6), clear blends of PMMA with a homopolymer of the same POSS macromer were obtained only at or below levels of 9% by weight of the POSS material. If optical clarity can be used as a measure of compatibility, the incorporation of the POSS material in a copolymer with MMA greatly expanded the amount of POSS material which can be incorporated in these blends. A series of MMA copolymers with increasing POSS content showed that these materials have superior

thermal stability compared to PMMA. Preliminary work on solvent-free compression molded blends suggests that these also have superior thermal stability relative to PMMA. Based upon the thermal characterization of these blends, more extensive characterization should be carried out.

SUGGESTIONS FOR FURTHER WORK

1. A series of solvent-free and nearly flaw free disks should be prepared by compression molding of the five materials prepared in this study.
2. For each of the five materials, disks 25mm in diameter and weighing 400 to 500mg should be prepared for Rheometry, and for X-Ray Diffraction. A third disk of this size should be prepared f or Dynamic Mechanical Analysis (DMA) and TMA.
3, For each of the five materials, disks 25mm in diameter and weighing 200mg should be prepared for measurement of oxygen permeability. These disks need to be crack-free or erroneous results will be obtained.
4. For the four blends (excluding the 80/20 blend), larger crack-free disks approximately 50mm in diameter and weighing 250mg should be prepared for measurement of mechanical properties on an Instron tester.
5. TGA should be run on solvent-free samples of all five materials.
6. Portions of solvent-free disks should be retained for careful DSC study using an oscillating DSC instrument which provides significantly better base-line control than the instruments used to date.

REFERENCES

1. Allcock, H. R. and Visscher, K. B. *Chem. Mater.* **1992**, *4*, 1182-1187.
2. Allcock, H. R. Visscher, K. B., and Manners, I. *Chem. Mater.* **1992**, *4*, 1188-1192.
3. Landry, C. J. T., Ferrar, W. T., Teegarden, D. M., and Coltrain, B. K. *Macromolecules* **1993**, *26*, 35-48.
4. Venugopal, G., Krause, S., and Wnek, G. *Chem. Mater.* **1992**, *4*, 1334-1343.
5. Lu, S., Pearce, E. M., and Kwei, T. K, *Polymer Engineering and Science* **1995**, *35*, 1113-1116
6. Noel, C. J. *Final Report, Summer Faculty Research Program* **1994**, Paper 24.
7. Lichtenhan, J.D. *Silsesquioxane-Based Polymers,* in The Polymeric Materials Encyclopedia, CRC Press, **1996**, *in press.*

SECONDARY ELECTRON EMISSION AT PLASMA-FACING SURFACES

Carlos A. Ordonez
Assistant Professor
Department of Physics

University of North Texas
Denton, TX 76203-0368

Final Report for:
Summer Faculty Research Program
Phillips Laboratory

August 1995

SECONDARY ELECTRON EMISSION AT PLASMA-FACING SURFACES

Carlos A. Ordonez
Assistant Professor
Department of Physics
University of North Texas

Abstract

The secondary electron emission coefficient is evaluated for electron impact on anode, cathode, and electrically-floating plasma-facing surfaces. Two energy and angle distribution functions for electron impact on a plasma-facing surface are derived and different relations for the secondary electron emission coefficient which functionally depend on energy and angle are integrated over the distributions. One integration is in closed form and provides a parametric expression for the secondary electron emission coefficient of a plasma-facing surface. The other integrations are carried out numerically. Evaluation of the secondary electron emission coefficient for a variety of commonly-used plasma-facing materials shows that its value is near or above unity over a significant range of plasma temperatures.

SECONDARY ELECTRON EMISSION AT PLASMA-FACING SURFACES

Carlos A. Ordonez

I. Introduction

In a large variety of plasma-producing devises, including those used for processing materials, electric space propulsion, electrical current control, radiation production, and fusion studies, plasma electrons bombard plasma-facing surfaces and cause the ejection of secondary electrons from the surfaces. This phenomenon is characterized by the secondary electron emission coefficient, δ, which equals the average number of electrons emitted from the surface per incident electron. Secondary electron emission from a plasma-facing surface has a significant effect on the value of the sheath potential.[1-4] This is true regardless of whether the surface carries a current or is electrically floating.[3] As a result, an accurate prediction of the secondary electron emission coefficient is needed both for a good theoretical understanding of particle and energy transport to plasma-facing surfaces[1,3,4] and for an adequate theoretical/computational description of sputtering and ion implantation at plasma-facing surfaces.[5,6] This provides the motivation for the work presented here. In Section II the secondary electron emission coefficient is evaluated for a surface which faces a Maxwellian plasma. Three different semi-empirical or empirical relations for the functional dependence of secondary electron emission on energy and angle of electrons incident on the surface are used for separate evaluations. The surface is considered to have a negative potential with respect to the plasma and is separated from the Maxwellian plasma by a plasma sheath which has a monotonically decreasing electric potential. In Section III, the secondary electron emission coefficient is evaluated numerically for an anode surface with a positive potential with respect to the Maxwellian plasma. The evaluation in Section III also applies to a surface with a negative potential bounded by a plasma sheath within which space-charge saturation occurs. Concluding remarks are given in Section IV.

II. Negative Surface Potential; No Space-Charge Saturation

The plasma configuration considered in this section is illustrated in Fig. 1. A plasma-facing surface which has a negative electric potential with respect to a Maxwellian plasma is considered. A plasma sheath, which exists between the Maxwellian plasma and the plasma-facing surface, is modeled as is done in fully-kinetic sheath theory[3,7] by considering a planar source of plasma particles which injects half-Maxwellian electrons and ions into the sheath. (The planar source and sheath represent both the presheath and the Debye sheath in a real system.) A Cartesian coordinate system is defined such that the x axis is normal to the plane of the surface. The planar source is located at $x = 0$ and is defined to have an electric potential of zero. Secondary electron emission at the plasma-facing surface is considered to be sufficiently small so that no space-charge saturation occurs within the plasma sheath. As a result, the electric potential within the sheath is a monotonically decreasing function of x. The surface can be electrically floating with respect to the Maxwellian plasma or it can be a cathode. The surface can also be an anode if the surface potential is negative-valued. (The value of surface potential must be positive relative to that of the floating potential to be an anode, however.) The four-dimensional phase space distribution function for plasma electrons within the sheath is a truncated Maxwellian given by Eq. (21) of Ref. [3]. At the sheath-surface interface, the velocity distribution function, with arbitrary normalization, is simply a half-Maxwellian:

$$f = e^{-\beta v^2} \; \Theta(v_x) \tag{1}$$

Here, $\beta = m/(2\ T)$, m is the electron mass, T is the electron temperature (in energy units) in the Maxwellian plasma, and Θ is the Heaviside step function. An energy and angle distribution is derived from Eq. (1) by carrying out a transformation from Cartesian coordinates to spherical coordinates in velocity space, assuming azimuthal symmetry about the surface normal (this eliminates the azimuthal coordinate), and making a second transformation to energy and angle coordinates. The resulting distribution function

is separable in energy and angle. With the two parts of the distribution normalized to unity, the distribution function is

$$f(E, \alpha) = f_E(E) \, f_\alpha(\alpha) \tag{2}$$

with

$$f_E(E) = 2 \sqrt{\frac{E}{\pi \, T^3}} \, e^{-E/T}$$

and

$$f_\alpha(\alpha) = sin\alpha$$

Here, E is the incident-electron kinetic energy and α is the angle of incidence with respect to the surface normal. The step function in Eq. (1) becomes, $\Theta(\sqrt{2 \, E/m} \, cos\alpha)$, which is satisfied with $0 < E < \infty$ and $0 < \alpha < \pi/2$. It should be pointed out that the effect of a magnetic field is not taken into account in deriving Eq. (2). Nevertheless, Eq. (2) should be suitable for a magnetized plasma provided either the magnetic field intersects the surface at normal incidents or the electron cyclotron radius is larger than a few Debye lengths.

The secondary electron emission coefficient has previously been considered to be separable in incident-electron energy and angle dependence: $\delta(E, \alpha) = \delta_E(E) \, \delta_\alpha(\alpha)$.[8] For the energy dependence at normal incidence, Kollath's relation is used. It is[8]

$$\delta_E(E) = \frac{(2.72)^2 \, \delta_m \, E}{E_m} \, e^{-2(E/E_m)^{1/2}} \tag{3}$$

where δ_m is the maximum value for δ at normal incidence and E_m is the energy at which the maximum value for δ occurs. The values for δ_m and E_m are constants which depend on the composition of the material surface. For $\delta_\alpha(\alpha)$, an angular dependence of the form,

$$\delta_\alpha(\alpha) = (sec\alpha)^g \tag{4}$$

with $g = 0.4$ has been reported.[9] Integrating $\delta(E, \alpha)$ over the energy and angle distribution, Eq. (2), provides the secondary electron emission coefficient in terms of the

plasma-electron temperature, T, and parameters determined by the surface material. This is possible in closed form giving the result,

$$\delta(T) = \int_0^\infty \int_0^{\pi/2} \delta(E, \alpha) \, f(E, \alpha) \, d\alpha \, dE$$

$$= \frac{2.1 \, \delta_m}{1 - g} \left[\sqrt{\pi} \, \tau \, (3 + 12 \, \tau + 4 \, \tau^2) \, e^\tau \, erfc\left(\sqrt{\tau}\right) - 2 \, \tau^{3/2} \, (5 + 2 \, \tau) \right] \tag{5}$$

where $\tau = T/E_m$ and $erfc$ is the complementary error function.

A semi-empirical relation for the secondary electron emission coefficient has been derived by Schwarz by considering a collision cascade process.[10] Combining Eqs. (22), (25a), and (25b) (with $n = \frac{4}{3}$) from Ref. [10] gives

$$\delta(E, \alpha) = \frac{2.6 \, \delta_m \, \epsilon^{2/3} \, (1 + 2 \, \epsilon \, cos\alpha)}{(1 + \epsilon \, cos\alpha)^3} \tag{6}$$

where $\epsilon = 0.72 \, E/E_m$. Integrating Eq. (6) over the angular part of the electron distribution function provides,

$$\delta(E) = \int_0^{\pi/2} \delta(E, \alpha) \, f_\alpha(\alpha) \, d\alpha = \frac{1.3 \, \delta_m \, \epsilon^{2/3} \, (2 + 3 \, \epsilon)}{(1 + \epsilon)^2} \tag{7}$$

Integration over the energy part of the distribution takes the form,

$$\delta(T) = \int_0^\infty \delta(E) \, f_E(E) \, dE = \frac{2.6 \, \delta_m}{\sqrt{\pi} \, r^3} \int_0^\infty \frac{\epsilon^{7/6} \, (2 + 3 \, \epsilon) \, e^{-\epsilon/r}}{(1 + \epsilon)^2} \, d\epsilon \tag{8}$$

where $r = 0.72 \, T/E_m$.

A relation for the secondary electron emission coefficient has also been presented by Vaughan.[11] The relation is empirical and is given by

$$\delta(E, \alpha) = \delta_m \, \left(1 + \frac{k_s \, \alpha^2}{2 \, \pi} \right) \, \left(v \, e^{1-v} \right)^k \, \Theta(v) \tag{9}$$

where

$$v = \frac{E - E_0}{E_m(1 + k_s \, \alpha^2/\pi) - E_0}$$

with $E_0 = 12.5$ eV and $k = 0.62$ for $v < 1$ or $k = 0.25$ for $v > 1$. The Heaviside step function is included to require $\delta(E < E_0, \alpha) = 0$ since it may otherwise be negative. The value of k_s depends on the surface roughness. The limits are $k_s \approx 0$ for a very rough surface and $k_s \approx 2$ for a very smooth surface. For the calculations presented here, $k_s = 1$ is used along with values for δ_m and E_m from Ref. [8]. Also, a continuous transition for k is used:[11]

$$k = 0.435 - \frac{0.37}{\pi} arctan \left[10 \; ln(v) \right] \tag{10}$$

In Fig. 2, calculations using Eq. (3) [with an angular dependence given by Eq. (4)], Eq. (6), and Eq. (9) are shown for a carbon surface (using $\delta_m = 1.0$ and $E_m = 300$ eV). The functional dependences on energy are found to be almost indistinguishable at normal incidence. However, at an angle of 1 radian, all three relations have different energy dependences. [It should be pointed out that, for the comparison in Fig. 2b, better agreement between Eq. (6) and Eq. (9) occurs if an unrealistically large value for the smoothness factor is used ($k_s = 3$). Also, one radian is chosen for the second comparison because it is the average angle of incidence for plasma electrons considered in this section.] As a result of the different energy dependences at large angles, Eq. (5), Eq. (8), and numerical integration of Eq. (9) over the energy and angle distribution, Eq. (2), give different predictions for the dependence of the secondary electron emission coefficient on plasma temperature. This is shown in Fig. 3a for a carbon surface. Also shown is an evaluation using Eq. (7) with $E = \frac{3}{2}T$ which is found not to be a suitable approximation for Eq. (8). In Fig. 3b, Eq. (5) is used to calculate the secondary electron emission coefficient for a variety of commonly-used plasma-facing surfaces. An important trend found in Fig. 3b is that a significant range of plasma temperatures exists over which plasma-facing materials have secondary electron emission coefficients near or above unity. (The exception to this is beryllium which has a secondary electron emission coefficient well below unity at all plasma temperatures.) With secondary electron emission coefficients near or above unity, space-charge saturation takes place in

plasma sheaths and simple relations for the sheath potential which neglect space-charge saturation effects are no longer valid.[4,9]

III. Positive Surface Potential or Space-Charge Saturation

In this section, the secondary electron emission coefficient is evaluated for an anode with a positive potential relative to the plasma it faces. The evaluation also applies to a negative-potential surface which has an adjacent plasma sheath within which space-charge saturation occurs. For these two cases, the electric potential profiles between the Maxwellian plasma and the surface are illustrated in Figs. 4a and 4b. Similar to Section II, a planar source of half-Maxwellian electrons is considered. The planar source injects electrons into a region with a monotonically increasing potential. For the anode, the planar source is located at the boundary of the Maxwellian plasma. For the sheath with space-charge saturation, the planar source is located at the potential minimum within the sheath (just before the increase in potential, ϕ, in Fig. 4b). Between the planar source and the surface, electrons are accelerated through the change in potential and strike the surface with a minimum velocity given by $v_m = \sqrt{2\,e\,\phi/m} = \sqrt{\psi/\beta}$ where e is the charge magnitude of the electron, ϕ is the difference in potential between the material surface and the planar source, and $\psi = e\,\phi/T$ is the normalized potential. The velocity distribution function for these electrons as they reach the surface is the same as the velocity distribution for half-Maxwellian ions after passing through a decrease in potential, Eq. (3) of Ref. [3]. With arbitrary normalization, the electron velocity distribution function at the surface is

$$f = e^{-\beta v^2}\,\Theta\left(v_x - v_m\right) \tag{11}$$

The corresponding energy and angle distribution is found by carrying out two transformations as described in Section II. Normalized to unity, the distribution is

$$f(E,\alpha) = 2\,\sqrt{\frac{E}{\pi\,T^3}}\,\frac{e^{-E/T}\,sin\alpha}{erfc(\sqrt{\psi}\,)}\,\Theta\left(E - T\,\psi\,sec^2\alpha\right) \tag{12}$$

$$30 - 8$$

This distribution must be integrated over energy first and then angle. Alternatively, integration over angle first and then energy is possible using the following product of Heaviside step functions in place of the single step function:

$$\Theta\left(E - T\,\psi\right)\,\Theta\left(arccos\sqrt{\frac{T\,\psi}{E}} - \alpha\right) \tag{13}$$

With Eq. (12), it is possible to integrate Eq. (6) over the angle variable and then simplify the remaining integral. [Eq. (3) with (4) can also be reduced to a single integral but the result is not shown in favor of Eq. (6) which is expected to have a more accurate angular dependence.] The result for integration of Eq. (6) is

$$\delta(T,\phi) = \frac{2.6\,\delta_m}{\sqrt{\pi}\,r^3\,erfc(\sqrt{\psi})} \int_{r\psi}^{\infty} \epsilon^{1/6}\,e^{-\epsilon/r}\left[\frac{3 + 4\,\sqrt{r\,\psi\,\epsilon}}{(1 + \sqrt{r\,\psi\,\epsilon}\,)^2} - \frac{3 + 4\,\epsilon}{(1 + \epsilon)^2}\right]\,d\epsilon \tag{14}$$

where, as before, $r = 0.72\,T/E_m$ and $\psi = e\,\phi/T$. As is expected, setting ψ in Eq. (14) to zero and rearranging reproduces Eq. (8). In Fig. 5, the dependence of the secondary electron emission coefficient on the normalized potential, given by Eq. (14), is shown for a fixed plasma temperature. The weak dependence on ψ can be explained as follows. As the potential increases, the average angle of incidence decreases while the average energy increases. This occurs because the accelerating potential increases v_x relative to $v_\perp = \sqrt{v_y^2 + v_z^2}$. A decrease in average angle tends to cause a decrease in δ while an increase in average energy tends to cause an increase in δ at the energies considered. The opposing effects cancel to a large extent and, as a result, δ shows a weak dependence on ψ.

IV. Conclusion

Three relations for the secondary electron emission coefficient, which functionally depend on energy and angle, have been integrated over two energy and angle distribution functions for electron impact on a plasma-facing surface. Only one of the integrations

is in closed form and provides a parametric expression [Eq. (5)]. With Eq. (5) it was found that many commonly-used plasma-facing surfaces have secondary electron emission coefficients near or above unity for plasma temperatures above ~50 eV. This indicates that space-charge saturation within plasma sheaths is a common occurrence which must be taken into account in order to accurately evaluate the sheath potential. Consequently, further work in sheath modeling which takes into account space-charge saturation effects appears warranted. Also, further work appears needed in secondary electron emission semi-empirical or empirical modeling to determine which existing (or newly developed) relation for the secondary electron emission coefficient, $\delta(E, \alpha)$, is most accurate at large angles of incidence.

References

[1] P. C. Stangeby in *Physics of Plasma-Wall Interactions in Controlled Fusion*, edited by D. E. Post and R. Behrisch (Plenum Press, New York, 1986) p. 41.

[2] R. N. Franklin and W. E. Han, Plasma Phys. Controlled Fusion **30**, 771 (1988).

[3] C. A. Ordonez, Phys. Fluids B **4**, 778 (1992).—

[4] L. A. Schwager, Phys. Fluids B **5**, 631 (1993); and references therein.

[5] C. A. Ordonez, W. D. Booth, R. Carrera, and M. E. Oakes, J. Nucl. Materials **185**, 130 (1991).

[6] C. A. Ordonez, W. D. Booth, R. Carrera, and M. E. Oakes, Rad. Effects and Defects in Solids **116**, 49 (1991).

[7] L. A. Schwager and C. K. Birdsall, Phys. Fluids B **2**, 1057 (1990).

[8] E. W. Thomas in *Data Compendium for Plasma-Surface Interactions*, (Nuclear Fusion, Special Issue, 1984) p. 94.

[9] J. M. Pedgley and G. M. McCracken, Plasma Phys. Control. Fusion **35**, 397 (1993).

[10] S. A. Schwarz, J. Appl. Phys. **68**, 2382 (1990).

[11] J. R. M. Vaughan, IEEE Trans. Electron Devices **36**, 1963 (1989).
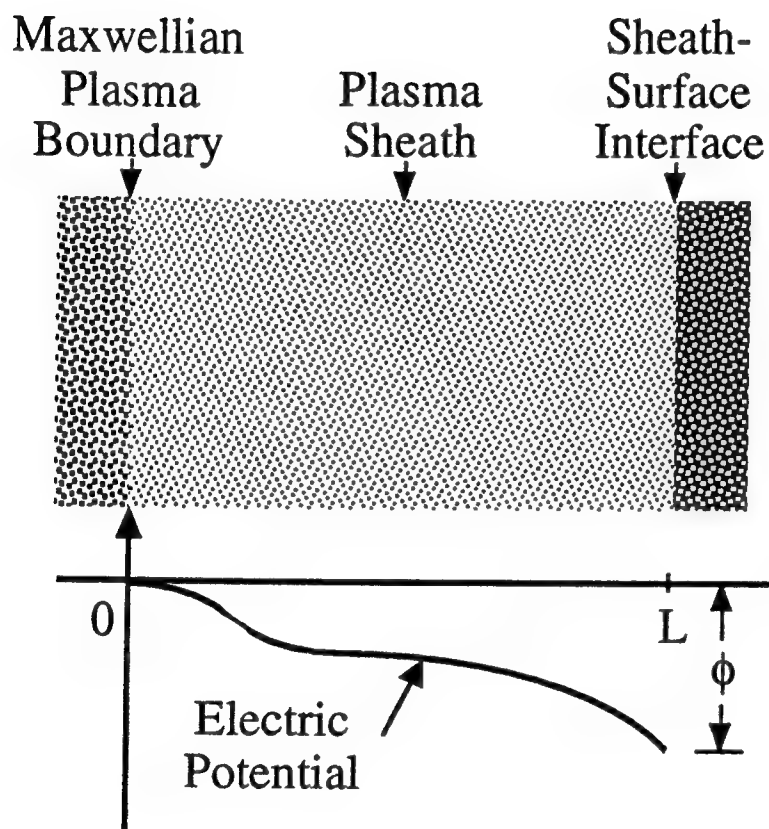
Fig. 1. The plasma configuration considered in Section II. A Maxwellian plasma injects half-Maxwellian electrons into a sheath region which has a monotonically decreasing electric potential.
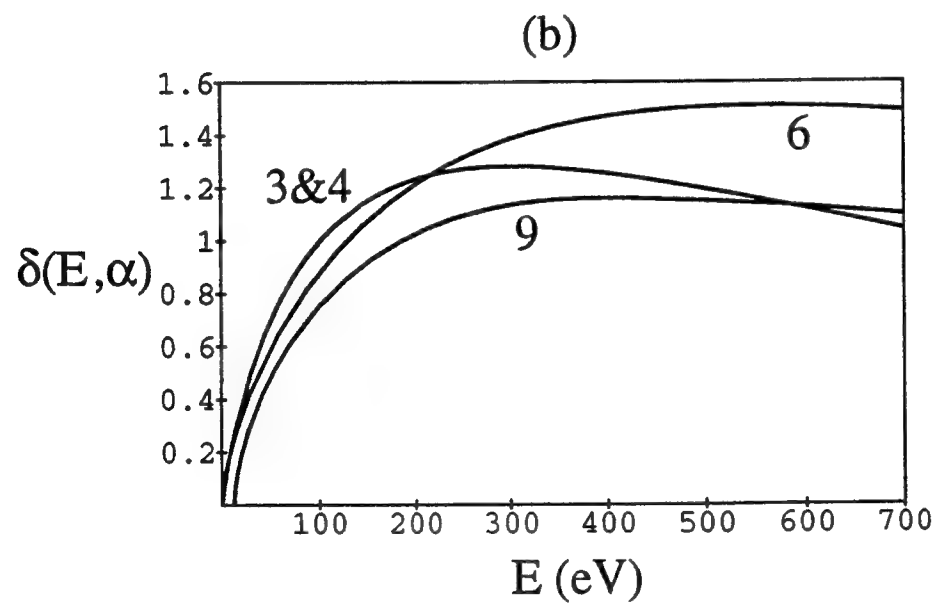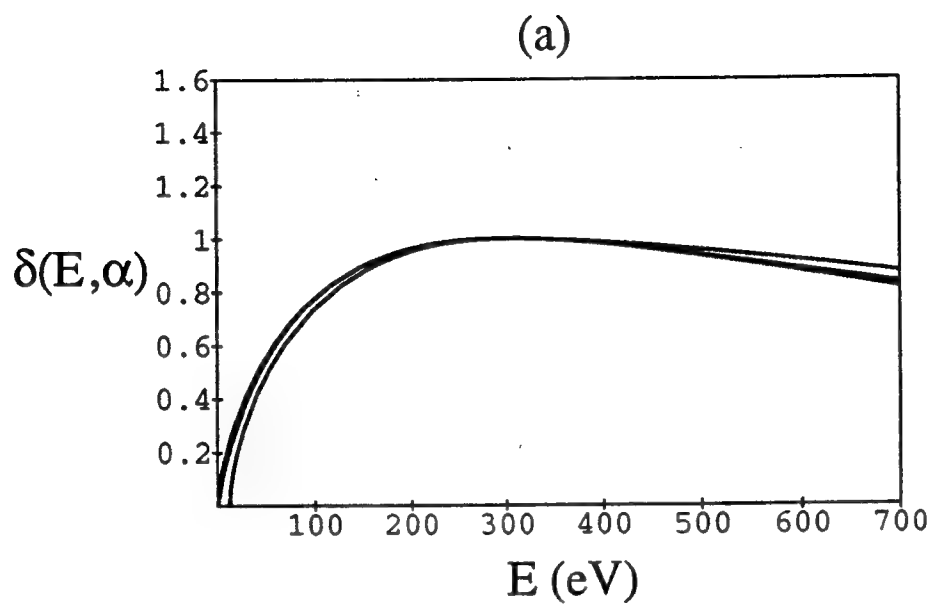
Fig. 2. Calculations of the secondary electron emission coefficient for a carbon surface using Eq. (3) with (4), Eq. (6), and Eq. (9). (a) $\alpha = 0$. (b) $\alpha = 1$ radian $= 57^o$.

(a)

(b)
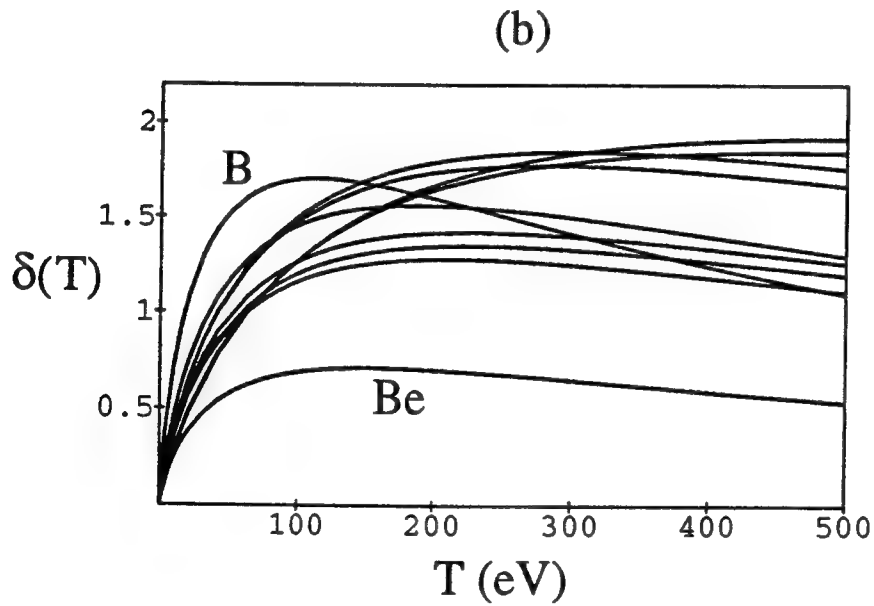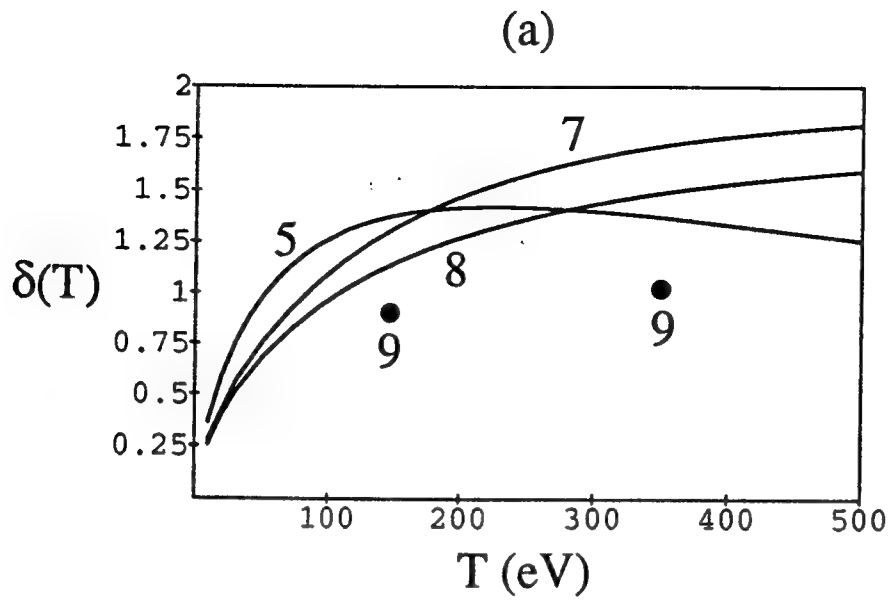
Fig. 3. Evaluation of the secondary electron emission coefficient. (a) For a carbon surface using Eq. (5), Eq. (7) with $E = \frac{3}{2}T$, Eq. (8), and Eq. (9) integrated over Eq. (2). (b) For Be, B, C, Al, Si, Ti, Fe, Cu, Mo, and W surfaces using Eq. (5).

30 – 14

(a)

(b)
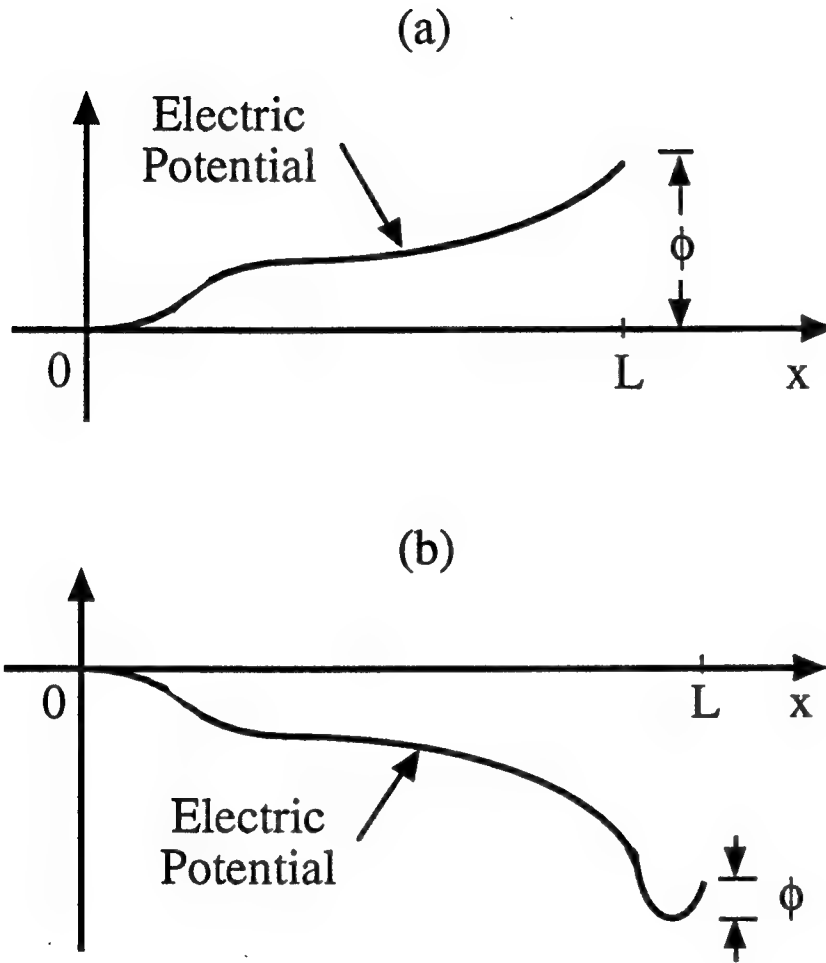
Fig. 4. Electric potential profiles considered in Section III between a Maxwellian plasma at $x \leq 0$ and a material surface at $x = L$. Shown are (a) an anode with a positive potential with respect to the Maxwellian plasma and (b) a surface having a negative potential with respect to the Maxwellian plasma but with space-charge saturation within the adjacent plasma sheath.
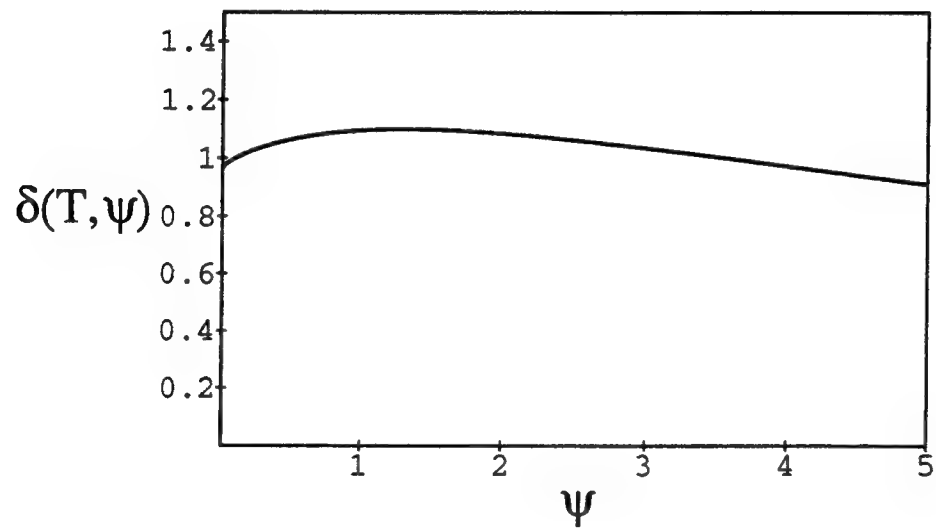
Fig. 5. Evaluation of the secondary electron emission coefficient for a carbon surface and a 100 eV plasma temperature using Eq. (14).

SET-UP FOR PLASMA INSTABILITY ANALYSIS OF KILOHERTZ ORDER WAVES ASSOCIATED WITH ELECTRON BEAM OPERATIONS ON STS 46

.

Michael J. Pangia
Assistant Professor
Department of Geology & Physics

Georgia Southwestern College
800 Wheatley Street
Americus, GA 31709-4693

SET-UP FOR PLASMA INSTABILITY ANALYSIS OF KILOHERTZ ORDER WAVES ASSOCIATED WITH ELECTRON BEAM OPERATIONS ON STS 46

Michael J. Pangia
Assistant Professor
Department of Geology & Physics
Georgia Southwestern College

Abstract

A certain set of electron data measured by the Space Particle Correlator Experiment (SPACE) during beam operations on STS 46 showed wave behavior in the 0.5 to 2 kHZ frequency range. This work derives a tractable equation from the full dispersion equation given in *Kindel and Kennel* [1971] to perform an instability analysis to account for the observations. It assumes that the waves can be explained by the two dominant plasma components, the O^+ ions, and the returning electrons. Further assumptions are that $k_\perp^2 A_e^2 < 2\Omega_e^2$ and that $\Omega_i < k_i A_i$, where A_e and A_i are the electron and ion thermal speeds, and Ω_e and Ω_i are the cyclotron frequencies.

Suggestions for further progress are made in the report.

SET-UP FOR PLASMA INSTABILITY ANALYSIS OF KILOHERTZ ORDER WAVES ASSOCIATED WITH ELECTRON BEAM OPERATIONS ON STS 46

Michael J. Pangia

I. Introduction

Electron wave behavior was directly measured by the Space Particle Correlator Experiment (SPACE), which was part of the Shuttle Potential and Return Electron Experiment (SPREE), an experiment package that was associated with the Tethered Satellite System 1 (TSS 1) which flew on STS 46 [*Oberhardt et al.*, 1994]. During active periods of the Fast Pulsed Electron Generator (FPEG), a certain set of SPREE/SPACE observations show strong electron modulations in the 0.5 to 2.0 kHz frequency range. It is the purpose of this study to develop a tractable equation to analyse reasonable parameter regimes in order to find the instability causing the waves.

II. Summary of Data

A detailed analysis of the data is done by *Gough et al.* [1995]. Here a brief list of their findings is given:

1)   Return electrons are associated with the beam operations constituting a return current. The return current has a velocity component parallel to the ambient magnetic field of about 3×10^6 m/s.

2)   The frequency range of the electron modulations are 0.5 to 2 kHz.

3)   The wave period of the modulations are proportional to the sine of the FPEG pitch angle.

III. Basic Equations

The wave mechanism is likely to be due to inverse Landau damping. The starting point

for the instability analysis would be the same as that studied by *Kindel and Kennel* [1971], who studied the properties of marginally unstable modes. The starting equation assumes isotropic Maxwellian velocity distributions and is the following

$$-k^2 = \sum_j \sum_{n=-\infty}^{\infty} \frac{\Gamma_n(\mu_j)}{\lambda_{Dj}^2} \left[1 + \frac{\omega - k_\parallel v_{dj}}{k_\parallel A_j} Z\left(\frac{\omega - k_\parallel v_{dj} + n\Omega_j}{k_\parallel A_j} \right) \right]$$

(1)

where

$\omega = \omega_r + i\gamma$ is the complex angular frequency,

k is the wave number,

$\perp$ and $\parallel$ refer to the directions perpendicular and parallel to the ambient magnetic field,

m_j, and T_j are the mass and temperature (in energy units) of the jth plasma species,

$A_j = (2T_j/m_j)^{1/2}$ is the thermal speed of species j,

v_{dj} is the parallel component of the drift velocity for species j,

Ω_j and ω_{pj} are the cyclotron and plasma angular frequencies for species j,

$\lambda_{Dj} = A_j/(2^{1/2} \omega_{pj})$ is the Debye length for species j, and

$\mu_j = k_\perp^2 A_j^2/(2\Omega_j^2)$.

The function Γ is defined as $\Gamma(\mu) = e^{-\mu}I_n(\mu)$, with I_n being the modified Bessel function of order n. The function $Z(\rho)$ is the plasma dispersion function, which for positive Im ρ is defined as

$$Z = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} du \, \frac{e^{-u^2}}{u - \rho} \quad \text{for } Im\rho > 0$$

(2)

and is defined for other values of Im ρ by analytic continuation; however, for an instability

31-4

analysis, only Im $\rho > 0$ is necessary. Note that (2) can also be written as

$$Z(\rho) = i \int_0^\infty ds \ e^{is\rho - s^2/4} \quad \text{for } Im\rho > 0 \tag{3}$$

Other useful relations for $Z(\rho)$ are

$$Z' = -2(1 + \rho Z) \tag{4}$$

where Z' is the derivative of Z with respect to ρ,

$$Z \approx -2\rho + i\sqrt{\pi}(1 - \rho^2) \quad \text{for } |\rho| < 1 \tag{5}$$

and, with Im $\rho > 0$,

$$Z \approx -\frac{1}{\rho}\left(1 + \frac{1}{2\rho^2}\right) \quad \text{for } |\rho| > 1 \tag{6}$$

The generating function for I_n is

$$e^{\frac{1}{2}\mu(t+1/t)} = \sum_{n=-\infty}^{\infty} t^k I_n(\mu) \tag{7}$$

which, among other things, shows that

$$\sum_{n=-\infty}^{\infty} \Gamma_n(\mu) = 1 \tag{8}$$

and with the help of (3) gives

$$\sum_{n=-\infty}^{\infty} \Gamma_n(\mu) Z(\rho + n\sigma) = i \int_0^\infty ds \ \exp\ (is\rho - s^2/4 - 2\mu \sin^2\frac{s\sigma}{2}) \tag{9}$$

where σ is real and Im $\rho > 0$.

It will be assumed that the instability can be explained by including just the dominant

plasma components, which are the returning electrons and the ambient O^+ ions. It is assumed that the chief role of the FPEG electrons is to induce the return current and heat the electrons. In solving (1), a reference frame must be chosen, of which two are logical, the rest frame of the ion distribution or the rest frame of the returning electron distribution. Figures 1 and 2 give diagrams of the two choices.
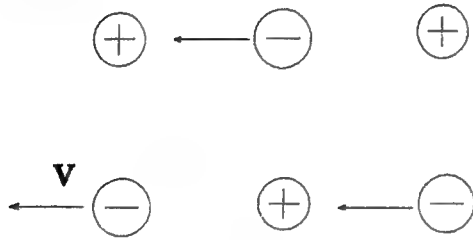
Fig. 1a. Particle motion in ion rest frame.
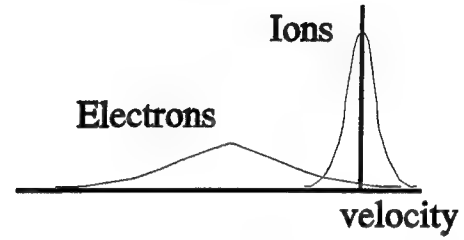
Fig. 1b. Distribution functions in ion rest frame.
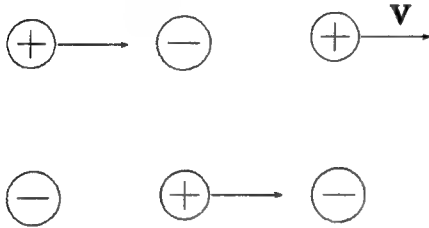
Fig. 2a. Particle motion in electron rest frame.
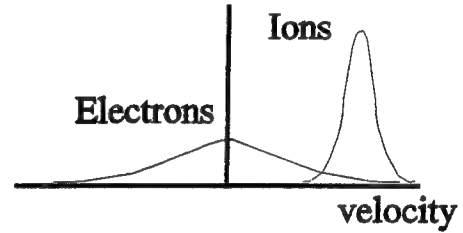
Fig. 2b. Distribution functions in electron rest frame.

By considering the form of the argument of the Z-function in (1), it is more appealing to choose the returning electron rest frame (Fig. 2). Then, at least, the cold temperature of the ions can lead to an approximation that will give simple expressions involving the product $k_{\parallel}v_d$ (v_d being the ion drift velocity). Therefore, in the rest frame of the returning electrons, (1) can be written (with use of (8))

$$-k^2\lambda_{De}^2 = 1 + \sum_{n=-\infty}^{\infty} \Gamma_n(\mu_e)\frac{\omega}{k_{\parallel}A_e}Z\left(\frac{\omega+n\Omega_e}{k_{\parallel}A_e}\right) + \frac{T_e}{T_i}\left[1 + \sum_{n=-\infty}^{\infty} \Gamma_n(\mu_i)\frac{\omega-k_{\parallel}v_d}{k_{\parallel}A_i}Z\left(\frac{\omega-k_{\parallel}v_d+n\Omega_i}{k_{\parallel}A_i}\right)\right]$$

(10)

The basic procedure would be to find the unstable roots of (10) within the parameter regime of the plasma. A marginal instability analysis (as was performed by *Kindel and Kennel* [1971]) would be insufficient. It is necessary to determine the most unstable mode in order to determine the dominant frequency.

Needless to say, (10) is too difficult to solve directly and either numerical or analytical approximations (or a combination of the two) would have to be used. This work will make analytic approximations to (10) to obtain a simpler equation.

IV. $\mu_e \ll 1$ and $\Omega_i \ll k_\parallel A_i$

Two reasonable approximations to make are $\mu_e \ll 1$ due to the large electron cyclotron frequency (about 1 MHz) and $\Omega_i \ll k_\parallel A_i$ due to the small ion cyclotron frequency (about 30 Hz). For $\mu \ll 1$, (9) up to the first order in μ becomes with the use of (3)

$$\sum_{n=-\infty}^{\infty} \Gamma_n(\mu) Z(\rho + n\sigma) \approx Z(\rho) + \frac{\mu}{2}\left(Z(\rho + \sigma) - 2Z(\rho) + Z(\rho - \sigma)\right) \text{ for } \mu \ll 1 \tag{11}$$

Setting $\mu = \mu_i$ and $\sigma = \Omega_i/k_\parallel A_i$ into (9), the approximation $\Omega_i \ll k_\parallel A_i$ leads to (for first order in σ)

$$\sum_{n=-\infty}^{\infty} \Gamma_n(\mu_i) Z(\rho + \frac{n\Omega_i}{k_\parallel A_i}) \approx \frac{k_\parallel}{k} Z(\frac{k_\parallel}{k}\rho) \text{ for } \Omega_i \ll k_\parallel A_i \tag{12}$$

Note that even though (12) was derived keeping to first order in σ, (12) is independent of Ω_i due to a cancellation with μ_i. However, (12) is not the same result gotten by simply neglecting Ω_i to begin with.

Using (4), (11), and (12), (10) becomes

$$-k^2 \lambda_{De}^2 = -\frac{1}{2} Z'\left(\frac{\omega}{k_\parallel A_e}\right) + \frac{1}{4}\frac{k_\perp}{k_\parallel}\frac{\omega}{\Omega_e}\frac{k_\parallel A_e}{\Omega_e}\left[Z\left(\frac{\omega + \Omega_e}{k_\parallel A_e}\right) - 2Z\left(\frac{\omega}{k_\parallel A_e}\right) + Z\left(\frac{\omega - \Omega_e}{k_\parallel A_e}\right)\right] - \frac{1}{2}\frac{T_e}{T_i} Z'\left(\frac{\omega - k_\parallel v_d}{k_\parallel A_i}\right) \tag{13}$$

V. Future Steps

Albeit (13) is subject to certain assumptions, it is more tractable for analysis than (10) and the assumptions are reasonable for the beam operations associated with STS 46. Therefore it is the suggested starting point for continued analysis of the problem.

There are some obvious next steps. One is to assume that the ions are cold enough so that (6) is valid. Another is to assume a strong instability so that $\gamma > k_i A_e$, which would also permit (6) to be used for the electron contribution. These can be tried together and would lead to algebraic expressions for the dispersion relation, which would have to be solved to find the most unstable frequency, its wave number and direction of propagation. Of course, the assumptions made along the way must the checked at the end to validate the results.

References

Gough, M. P. Gough, D. A. Hardy, M. R. Oberhardt, W. J. Burke, L. C. Gentile, B. McNeil, K. Bounar, D. C. Thompson, and W. J. Raitt,Correlator measurements of kHz wave-particle interactions during electron beam operations on STS 46, *submitted to J. Geophys. Res.*, 1995.

Kindel, J. M., and C. F. Kennel, Topside Current Instabilities, *J. Geophys. Res., 76*, 3055, 1971.

Oberhardt, M. R., D. A. Hardy, W. E. Slutter, J. O. McGarity, D. J. Sperry, A. W. Everest III, A. C. Huber, J. A. Pantazis, and M. P. Gough, The shuttle potential and return electron experiment (SPREE), *Il Nuovo Cimento, 17*, 67, 1994.

EXPLORATORY STUDIES OF AN IMAGE
MOTION ANALYSIS SYSTEM

Ronald M. Pickett
Professor
Department of Psychology

and

Institute for Visualization and Perception Research
University of Massachusetts Lowell
One University Avenue
Lowell, MA 01854

Final Report for:
Summer Faculty Research Program
Phillips Laboratory

August 1995

EXPLORATORY STUDIES OF AN IMAGE MOTION ANALYSIS SYSTEM

Ronald M. Pickett
Professor
Department of Psychology

and

Institute for Visualization and Perception Research
University of Massachusetts Lowell
One University Avenue
Lowell, MA 01854

Abstract

Tests were conducted of a system for analyzing the motion of non-rigid structures in time-lapse images. Tests with well specified artificial images showed that the system operates as intended. Tests with real images of depletion patterns in allsky images of the ionosphere, taken with a fisheye lens, show that when operated in a high-threshold mode the system produces motion data consistent with what would be expected. In a low threshold mode, the system produces reliable findings of motions that would not be expected. In pursuit of an explanation, it was determined that high frequency components of the images contribute little if anything to the motion findings. This would be consistent with the possibility that the aberrant findings are due to optical or geometrical distortions in the allsky image as the pattern moves past the camera, but a final explanation has not been established.

EXPLORATORY STUDIES OF AN IMAGE MOTION ANALYSIS SYSTEM

Ronald M. Pickett
Professor
Department of Psychology

and

Institute for Visualization and Perception Research
University of Massachusetts Lowell
One University Avenue
Lowell, MA 01854

Introduction

A basic component of research on the composition, structure and dynamics of the ionosphere is the recording and analysis of ground-based images of the night sky. These allsky images, taken with a fish eye lens at different wave lengths, provide a basis for identifying fluctuations in the composition and excitation of the ionosphere over space and time, which can then be related to other geophysical variables.

The challenge in analyzing these images -- to detect structures, to measure those structures and track them over time -- is to deal with their non-rigidity, with the fluidity of the medium in which they emerge and evolve. Several systems for analysis of the motion of non-rigid structures have been developed in other settings (Barron, et al, 1992). One, developed specifically for application in the present ionospheric research context (Gonthier, 1994), was the object of study here. The aim was to conduct a variety of exercises with this system -- to check on its basic technique in analyses with a variety of artificial test patterns, and then to explore its sensitivity and validity in analyses of some very simple structures and motions in real allsky images.

Problem Statement

The basic requirement in analyzing the motion of objects in time-lapse images is to find the object of interest in the image taken at time 1, find the same object in the image taken at time 2 and then measure any differences in position relative to the fixed frame of the images. When the object has a rigid shape, it is a relatively simple computational task to take a template of the object's outline in image 1, line it up with the object in image 2 and then measure the difference in position of the template between the two images. However, when the object can change in shape between the two images, the problem is greatly complicated. The general solution is to provide for some more

flexible way for the object to find itself in the second image, one that does not require an all or none fit to a fixed template.

The particular approach we have taken is to provide for every fragment, pixel or larger $n \times n$ block of pixels, in image 1 to find itself in image 2. For each fragment in image 1, we determine the gray scale values in the $w \times w$ window of fragments surrounding it. We then go to the other image, and starting at the locus where the fragment would be if it had not moved, we determine the gray scale values in the same sized surrounding window and compute an index of correlation. We then move the window around in image 2 to each of the neighboring fragments in a search area surrounding the no-movement locus and compute the correlation at each locus. The system then chooses, as the locus that the fragment moved to in the second image, the one that yielded the highest correlation, or, where greater stability is required, the center of gravity of the k loci yielding the highest correlations. For further details, see Gonthier (1994).

The system provides for the scientist to set several parameters, including two that determine the fineness of resolution of the analysis -- the fragment size, and the correlation window size -- and one that determines the search distance, how far the search will extend up, down, left and right from the no-movement locus.

The system then provides for various ways of probing, visualizing and analyzing the resulting data on where each fragment moved to in the second image. Two forms of visualization are provided. One indicates the direction of movement by the color that the fragment is painted in a color overlay on the first image. A distinctly different color for each of six cardinal directions is used, e.g., red for east, blue for west, etc.. In the other form, both the direction and the distance of movement is indicated by a colored arrow at each point. For each of these forms of display, the scientist can set the threshold value of the correlation to be exceeded for the fragment to be painted in the overlay. Thus, the scientist can see at glance, in the color overlay, how the individual fragments comprising the image moved between frames, and do that, for different levels of confidence in the data, by adjusting the correlation threshold.

The system also provides a basis for analyses of the motion data -- e.g. for computing the centroid of where the fragments comprising any selected object or region in image 1 moved to in image 2. Perhaps of most promise is the potential for using the fragment-by-fragment data to draw flow lines -- to show how selected fragments have moved from image to image over an extended sequence.

This motion analysis system provides a potentially powerful tool for a point by point visualization of motion in the image, for quantitative analyses of the motion data, and for depicting flow lines.

However, before applying it routinely in research a certain amount of testing and exploratory exercising of the system has been needed, and that was the primary goal of the present project.

Methods

We conducted two sets of tests. One was with test patterns of well specified artificial data -- to verify that the system was performing exactly as intended. The other was with real allsky images containing relatively simple structures and components of motion -- to determine how well the system could perform in bringing out the expected structures and motion.

Studies with artificial test patterns. We performed two kinds of studies with artificial test patterns. The first employed test patterns containing no macrostructure, just a consistent background of random microstructure. The gray scale value of each pixel in each image was selected independently from a gaussian distribution. The aim of the tests on these purely random images was to provide a fundamental statistical test of the operation of the system. Since the images in any pairing are totally independent, correlations between any window in one and any window in the other are strictly due to chance. Moreover, the number of such chance correlations that will exceed any particular value is a very predictable joint function of the correlation window size and the size of the search area. Counts of the number of correlations exceeding selected specific values, thus, provides a basis for a fundamental check that the system is, or is not, generating correlations at the predicted chance frequencies.

The other set of studies with artificial test patterns employed pairs of independent gaussian images as above, but with simple structures imposed on them. We looked at the system's ability to find where objects of different sizes and shapes moved from image 1 to image 2. How we imposed these objects varied from test to test. For example, in one set of tests, we selected a vertical rectangular region in image 1 and added a constant value to each pixel in it. We then repeated that operation in image 2 for a rectangular region of the same dimensions but in a position further to the right. We then ran the motion analysis to see how well the system worked to find a region defined in this way -- just on the basis of a mean difference between the region and its background. In other tests, we defined and imposed the region differently. We left the region in image 1 as it was originally generated with no mean difference between it and the back-ground. We then inserted a replica of it, pixel for pixel, in image 2 in the same right shifted position as in the first tests. This provided a basis for determining how well the analyzer worked to find an object defined only by its persistent microstructure. In other studies we looked at the combined effects of both

microstructural information and overall contrast information on the system's ability to detect and track the imposed object.

Studies with real allsky images. We used for these tests a series of images of ionosphere depletion patterns taken from ground-based locations near the equator. These patterns have a very simple characteristic structure -- alternating broad bands of high and low image intensity running north south, with a slow drift of the whole pattern from west to east.

We first conducted tests to determine what settings of the several adjustable parameters of the system brought out best the structure and motion that was evident from just looking at the images. We then conducted further studies aimed at understanding a perplexing departure that we found from what we expected. What that departure was and the various possible explanations we considered are presented below in the results section. Suffice it to say here that a basic question of interest was the source of information on which the system was finding reliable correlations. Was it solely or mainly in the large shifts in mean intensity between the bands? Or, was there also some reliable information in the microstructure? Was the microstructure persistent enough overtime for the system to find correlations in its fine details, not just in the mean shifts in intensity at the boundaries between the bands? In one test we looked at the effect on reliability of the motion data as the time lapse between the images was increased. If the correlations were dependent in a significant way on the microstructure, they would be expected to weaken drastically with increasing separation in time. Other tests were conducted with filtered images. If the correlations depended on information in the microstructure, a high-pass filtered image should produce much less reliable motion data than an unfiltered or a low-pass filtered image would.

Results

Tests with artificial images. Our studies with the independent gaussian background images produced correlations exceeding selected values at statistically predicted frequencies, providing strong evidence that the basic correlation technique was working as intended. An additional product of this activity are data for guidelines to user scientists on where to set the correlation threshold, for different choices of window size and search area, to be sure that the resulting motion data area reliable.

The studies with simple objects imposed on independent gaussian backgrounds showed that the motion analyzer can find and track objects in the images solely on a difference in mean gray scale level between the object and the background. Visualizations of the motion data from these

analyses, confirm what would be predicted -- that the object is detected in this situation on information only in the boundary. The other studies, where the object has no mean difference, showed that the system can find and track the object very sensitively solely on information in its microstructure, provided, as we ensured, that its microstructure persists into the second image. No difference in mean brightness of the object with its background is required for it to be found. One important implication of this is that the motion analyzer has the capability to detect and track a region that might not be evident in visual analyses of the images.

Tests with real allsky images. For these tests we depended mainly on the color overlays to interpret the findings. We expected that the motion analyzer would find a general overall pattern of motion from west to east -- i.e. that the predominant coloring of the visualization overlay would be red, and probably a vertically striated red, due to picking up the motion information first and perhaps mainly in the edges of the bands. This expectation came from our assumption that the correlations would be strongest at those places in the image where there is a steep gradient that can support a strong correlation even if there is no reliable information in the microstructure. For each test, we started our visual analyses of the color overlay with the threshold set very high. Then we lowered the threshold to see where and what color the fragments were that came in first. No matching fragments were found at very higher threshold settings, but as the threshold was lowered fragments began to be painted in -- mostly red (indicating west to east motion) and mostly at the edges of the bands as expected. As the threshold was lowered still further those strips of red broadened. Thus at a high threshold the analyzer was found to be doing what would be predicted. However, as the threshold was lowered further we began to see a filling in of blue fragments indicating that the system was finding fragments moving westward. This was and still is a perplexing finding. We have no reason to expect such a countervailing component of motion -- either on the basis of visual inspection of the specific images or from theory about the underlying physical structures and how they move.

The possibility that there is some defect in the motion analyzer always has to be considered, but careful reviews of the system and our tests with artificial patterns have yet to reveal an explanation from that quarter. Another possibility is that these backward motions are an artifact of geometrical or optical distortions in positioning of the boundaries, or in the boundary gradient profile as the pattern moves past the camera. The fact is that these real patterns do not slide linearly past the camera. Nor are they of a simple planar structure like our artificial test patterns simulate. These images are of a volumetric structure. The depletions are holes or bubbles in a thick layer of the ionosphere. This adds aspects of three dimensionality to consider in the possible distortions in the pattern as it moves past the camera.

Conclusions

The system stands up to a variety of tests with well specified artificial test patterns. The system also finds in the real images what it would be expected to find, provided it is operated at high threshold settings. At lower threshold settings, however, it finds a weak but reliable component of motion in the images that would not be predicted, either from visual inspection of the particular test image we used, or from consideration of what should occur in the general case.

In pursuit of an explanation of this perplexing departure from the expected, we conducted a variety of studies to determine how much the high frequency components of the image contribute to the motion data. These tests show consistently that the high frequency components contribute little or nothing to the motion data. This finding does not provide an explanation in itself, but it is consistent with the possibility that the perplexing component of motion is an artifact of some geometrical or optical distortions that occur as the depletion pattern moves across the sky and past the fisheye lens.

This motion analyzer remains a very promising tool for analyzing ionospheric images, but without an explanation for this perplexing finding, its application in this or other ionospheric research applications must be suspect. Further study is needed.

References

Barron, J.L., D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt, *Performance of Optical Flow Techniques,* IEEE, March 1992, pp 236-242.

Gonthier, D.L., *A Tool for the Analysis of Motion of Non-Rigid Structures,* M.S. Dissertation, (Computer Science) University of Massachusetts Lowell, 1994.

ON THE FLUID DYNAMICS OF SUPERCRITICAL DROPLET BREAKUP: AN ANALOG WITH MISCIBLE FLUIDS

Dimos Poulikakos
Professor

and

Chad J. Mitts
Graduate Student

Department of Mechanical Engineering
The University of Illinois at Chicago
842 West Taylor Street
Chicago, Illinois 60607-7022

ON THE FLUID DYNAMICS OF SUPERCRITICAL DROPLET BREAKUP: AN ANALOG WITH MISCIBLE FLUIDS

Dimos Poulikakos and Chad Mitts
Department of Mechanical Engineering
The University of Illinois at Chicago
842 W. Taylor Street, Chicago IL 60607

Summary

Atomization of liquid propellants and oxidants in rocket engines occurs in a high temperature and pressure environment, at conditions often exceeding the critical temperatures and pressures of the working liquids. It is therefore important that an extensive knowledge base exists on a host of relevant phenomena, exemplified by atomization in sprays undergoing transcritical transition. Such a knowledge base is currently very limited. The work reported herein aims at the above deficiency and reports preliminary results on droplet deformation and breakup in a miscible fluid analog of actual transcritical and supercritical droplet breakup. More specifically, this analog consists of a liquid droplet falling freely into a reservoir of another liquid in which the droplet is miscible. Two combinations of such liquids were examined in the context of the present study: sucrose solution droplets falling in a water reservoir, and petroleum oil droplets falling in a pentane reservoir. The study reveals a host of distinct breakup regimes for the falling droplet depending on the values of the droplet Reynolds number and the viscosity ratio of the miscible liquids. In addition, the importance of solubility becomes apparent and it is recommended that its effect is investigated in a future study.

1. Introduction

As the operating pressures and temperatures in contemporary liquid rocket engines and gas turbines increase towards supercritical conditions, an understanding of the basic mechanisms and regimes of droplet atomization under these conditions becomes a crucial component of cost effective, efficient design. Currently, the majority of the design codes utilize models for fuel atomization that were developed for subcritical conditions. These models do not include the major features of supercritical atomization phenomena such as the vanishing of surface tension, similar properties between the gas and liquid phases, the high solubility of the gas phase in the liquid phase which can lead to a mixture with an increased critical pressure, and property singularities near the critical point. To begin addressing this problem, the behavior of a single liquid droplet under transcritical and supercritical conditions is examined. The knowledge accumulated is a building block toward the understanding of the more complex sprays and jets present in actual injectors in rocket propulsion.

The main focus of the research was the development of a zero surface tension analog of the real droplet deformation and breakup process, through the use of miscible fluids. An additional scope of the work was the

characterization of a novel piezoelectric LOX droplet generator, built at Phillips laboratory. The goal of the miscible fluid supercritical analog work is to characterize the breakup mechanisms of a zero surface tension droplet over a wide range of viscosity ratios and Reynolds numbers. By establishing similitude with the real supercritical case through the proper selection of miscible fluids, it is hoped that the mechanisms of supercritical droplet deformation and breakup can be captured. The investigation of miscible fluids as a supercritical droplet simulation was initiated at Phillips laboratory [1]. The goal of studying the LOX droplet generator is to gain a better understanding of the stable operation of the device so that its use (at a later time) in the study of supercritical phenomena is reliable and devoid of unexpected complexities. To this end, two experiments were established, one which allowed the evolution of a zero surface tension droplet to be observed over a wide range of viscosity ratios and Reynolds numbers, and another which was capable of producing and recording the images of a continuous stream of LOX droplets at atmospheric conditions.

2.   Background

Supercritical Droplet Breakup Analog

Traditionally, the deformation and breakup of liquid particles (droplets) has been characterized by the density and the viscosity of the particle and the fluid surrounding, the particle effective diameter, the velocity of the particle relative to the surrounding fluid, and the particle surface tension. In addition, the acceleration of the particle has been included as a parameter when it is important. These properties are grouped into the following nondimensional numbers [1]: the density ratio (DR), the viscosity ratio (VR), the Weber number (We), the Eotvos number (Eo), and the Reynolds number (Re). In many situations, the Ohnesorage number (Oh) is commonly used in place of Re in describing the breakup regimes of subcritical droplets. All of these numbers are defined below.

$$DR = \rho_p / \rho \tag{1}$$

$$VR = \mu_p / \mu \tag{2}$$

$$We = \rho D V^2 / \sigma \tag{3}$$

$$Eo = a\rho_p D^2 / \sigma \tag{4}$$

$$Re = \rho V D / \mu_p \tag{5}$$

$$Oh = \mu / (\rho_p D \sigma)^{1/2} \tag{6}$$

The symbols ρ, σ, μ, D, V and a stand for density, surface tension, viscosity, particle diameter, particle velocity and particle acceleration respectively. The subscript p denotes the particle (droplet). Under transcritical and supercritical conditions, the surface tension approaches zero, thus We, Eo, and Oh all approach infinity and are no longer a useful means of characterizing the droplet behavior. In addition, densities of the liquid phase and the gas

phase under supercritical conditions approach the same order of magnitude, thus the density ratio also becomes uninformative. This suggests that the evolution of a liquid particle in supercritical conditions can be solely described based on the viscosity ratio and the Reynolds number [1]. This appears reasonable since these two numbers describe the influence of the two primary forces that govern particle deformation and breakup in the absence of surface tension, the inertial forces which tend to induce breakup, and the viscous forces which tend to resist particle breakup.

To determine the accuracy with which the deformation and breakup of a zero surface tension particle can be described through the combination of the viscosity ratio and the Reynolds number, a simple experimental setup was developed that utilizes two miscible fluids to simulate the zero surface tension condition of a supercritical environment. The premise of this model is that the surface tension, which is caused by an asymmetry of molecular forces at the interface of two fluids [2], will vanish as the distinct fluid interface disappears due to the effect of the fluids dissolving into each other. While this method does not replicate the thermodynamic singularities or thermal transport properties relevant in supercritical droplet study, it should still capture the purely kinematic nature of particle deformation and breakup. Furthermore, the miscibility of the two fluids is similar to the high solubility of the gas phase into the liquid phase.

Piezoelectric LOX Droplet Generator

One of the primary obstacles in the study of supercritical oxidants and fuels is the stable generation of a monodispersed stream of droplets. To meet this challenge, a piezoelectric droplet generator was developed at Phillips laboratory that is capable of producing LOX droplets by condensing gaseous oxygen (GOX). It utilizes a concentric tube heat exchanger arrangement, with a continuous flow of LN2 serving as the cooling fluid for the GOX. The advantage of this device is two fold. First, LOX is an extremely dangerous substance. Concentrated amounts of oxygen can easily oxidize many items. By condensing GOX instead of feeding the generator directly from a LOX reservoir, the amount of LOX present in the laboratory is greatly reduced yielding a much safer practice. Second, gaseous oxygen and LN2 are readily available in most laboratory environments, whereas LOX is more difficult to obtain and maintain.

Once the stream of oxygen is condensed in the heat exchanger, it passes through a cylindrical piezoceramic tube that is pulsed by a function generator. With each pulse, the tube contracts and relaxes, thus inducing perturbations in the LOX stream. In principle, the optimum frequency of perturbation needed to induce Rayleigh breakup was determined by Weber [3] as a function of the jet diameter and velocity and it is given below in equation (7).

$$\frac{1}{f} = \frac{\sqrt{2}\pi d}{v}\left(1 + \frac{3\mu_L}{\sqrt{\rho_L \sigma d}}\right)^{0.5}$$ (7)

The symbols f, v, ρ, μ, and d refer to the optimum frequency of oscillation, the mean jet velocity, the fluid density, viscosity, and the diameter of the jet respectively. In this experiment, only one orifice diameter was available, therefore, the jet diameter remains relatively constant. Thus, a change in the piezo oscillation frequency must be balanced by the appropriate change in the mass flow rate of the oxygen in order to excite the optimum instability leading to a Rayleigh breakup mode. The effect on the droplet stream due to changes in the frequency (for $Re_{jet} <$ 10^3) should be reflected mainly in a change in the velocity of the droplets. The diameter of the droplets will largely be determined by the orifice diameter.

3. Experimental Setup and Procedures

Supercritical Droplet Breakup Analog

A significant advantage to the study of zero surface tension particles using the miscible fluid analog is the simple and inexpensive experimental setup that is required (Fig. 1). The experimental setup consisted of a transparent, Plexiglas test section of four inches square in cross section (normal to the direction of motion of the particle) and twelve inches in height. Above the test section, a two mm diameter flat tip tube and syringe were mounted on a calibrated rail such that the height of the tube above the surface of the fluid in the test section, denoted as h in Figure 1b, could be easily adjusted and measured. The particle fluid would collect at the end (tip) of the tube until the force of gravity overcame the surface tension and caused the droplet to release and fall towards the test section. A strobe light was mounted behind the test section, and it was used in combination with a Stanford Computer Optics 4 Quik 05 CCD camera with an image intensifier to record the evolution of the particle as it traveled through the test section. The image acquisition system consisted of a Data Translation QuickCapture Frame Grabber board installed in a Macintosh 7200/80 Power PC (Fig. 1). The image acquisition / analysis software used was NIH Image V1.57.

Two different sets of miscible fluids were used in the experiments: the first was a sucrose droplet falling into water, and the second was a petroleum oil droplet in pentane. In each of the cases, the substance with the greater density was used as the particle fluid so that the it would fall freely through the test section. Measured quantities of the ambient fluid were added to the particle fluid to change its viscosity. The viscosity and density were then measured as a function of the volume percentage of particle fluid to ambient fluid. The viscosity measurements were taken using a Cannon-Feaske viscometer according to the standard guidelines established in ASTM D446-74. In addition, for both the oil/pentane mixture and the sucrose water/mixture, the initial diameter of the particle formed from the tube was measured using the CCD camera and the NIH image analysis software.

For both the sucrose/water and the oil/pentane combinations, particle deformation and breakup images were digitally recorded for particle fluid volume percentage mixtures ranging from 100% to 20%, and for values of h ranging between one cm and six cm. The varying mixtures and heights provided a viscosity ratio range from 2.2 to

843 and a Reynolds number range from 2 to 1760 for the sucrose and water combination. For the oil and pentane experiments, the viscosity ratio ranged between 2.9 and 835 and the Reynolds number between 6.2 and 2,100 . The image data were used to characterize the breakup regimes for the different conditions, and to quantify various dimensional aspects of the particle.

Piezoelectric LOX Droplet Generator

A droplet generator was constructed and mounted to a stand such that the stream of LOX droplets was directed downward. A flow control valve on the gaseous oxygen feed line allowed the flow rate to be easily adjusted. A function generator and an amplifier were used to produce a square wave of 10 μs duration with a variable frequency and amplitude to drive the piezoelectric ceramic device. Reference 4 contains a detailed description of the droplet generator. A schematic of the experimental setup and droplet generator is given in Figure 2. Quantitative measurements of the droplet diameters were made using a Greenfield Instruments Model 700 Spray Analyzer. This device consisted of a CCD camera and a strobe light combination that sent shadowgraph images of the droplets to a Gateway 2000 PC for analysis using the Greenfield Instruments software. The frequency of the piezo oscillations was varied between 200 Hz and 2000 Hz. At each different frequency, the flow rate of the oxygen was adjusted to attain a Rayleigh breakup mode, and the diameter data was collected.

4. Results and Discussion

Supercritical Droplet Breakup Analog

The breakup regimes for the sucrose/water combination are shown in Figure 3 as a function of the viscosity ratio and the Reynolds number. The four distinct regions identified were viscous globules, viscous bag, vortex transition, and vortex ring behavior. The oil/pentane data are presented in Figure 4 again as a function of the viscosity ratio and the Reynolds number. The same basic breakup modes, viscous dominated and vortex ring, were identified for both combinations of fluids, however, the transition from the viscous dominated regime to the purely vortex behavior is much more distinct in the case of the oil/pentane combination. The transition region also occurs over a considerably smaller range of Reynolds numbers for the oil/pentane case, spanning values of 60 to 90, as compared a Reynolds number range of 85 to 500 for the sucrose/water experiments. The noted differences between the two combinations of miscible fluids under similar viscosity ratios and Reynolds numbers indicate that an additional mechanism, the solubility of the two fluids, is affecting the breakup of the particles beyond eliminating the surface tension. Therefore, the complete characterization of the deformation and breakup of a zero surface tension particle must include the solubility in addition to the viscosity ratio and the Reynolds number.

In the absence of surface tension, the primary force that resists breakup of the particle is the internal viscous force. The effect of solubility is that it tends to alter the viscosity of the particle fluid as the ambient fluid is

dissolved into it. Thus, the ability of the particle to resist deformation from the inertial forces is also altered. To demonstrate the significant effect that solubility can have, the distance from the free surface of the test section to the location that primary breakup of the particle occurs is nondimensionalized with respect to the particle initial diameter and plotted as a function of the viscosity ratio in Figure 5. The disparity of the distance to breakup between the oil/pentane and sucrose/water data is roughly an order of magnitude different for viscosity ratios of 400 and greater. At smaller viscosity ratios, the solubility rates are lowered as the concentration gradient between the particle and the ambient fluid is reduced. As a result of the smaller concentration gradients and the relatively weaker internal viscous forces at the lower viscosity ratios, the particle deformation is dominated by the inertial forces and the development of the particle occurs at a fast enough rate, such that the solubility is not a significant factor, thus, the data for the two different fluid combinations begin to converge.

Figure 6 illustrates the viscous globule breakup mode for the sucrose/water combination at a VR of 843 and a Re of 4.8, and it is a typical example for VR of 400 and greater. As the initially spherical particle enters the test section fluid, the pressure distribution around the particle will exhibit a maximum at the stagnation point near the center of the particle leading edge. The mass at the trailing edge of the particle begins to redistribute itself around the periphery of the particle in response to the leading edge mass that has been slowed down at the stagnation point. This redistribution creates the flat disc of particle fluid shown in Figure 6a. Corresponding to the initial stagnation point analysis, a minimum pressure would exist along a circumferential region slightly aft of the leading edge surface. However, due to the relatively high viscosity ratio of this case, the internal viscous force is great enough such that this suction has little effect on the particle development within the viscous globule regime. Due to the internal flow of mass towards the particle periphery, the center of the particle thins until it is no longer capable of resisting the shear stresses created by the pressure difference between the forward and aft sides of the particle, thus it consequently breaks. As a result of the new azimuthal distribution of mass, the maximum pressure shifts from the original stagnation point to the leading edge of the of the circumferential ring that has developed. The pressure distribution around the ring inherently exhibits variations, thus the ring becomes unstable and it breaks into separate, smaller viscous globules as demonstrated in Figure 6c. The smaller globules behave as the initial particle and generally, they would continue to break into smaller components at least one more time.

Figure 7 is the breakup of an oil droplet as it travels through pentane. The initial VR and Re are 835 and 2 respectively, which is similar to the sucrose particle conditions of Figure 6. The initial deformation is very similar between the two cases, with the initially spherical particle flattening out into a disc shape in response to the stagnation point at the leading edge. However, the pentane appears to dissolve into the oil at a faster rate than the water into the sucrose. This results in a diminished internal viscous force and as indicated in Figure 5, the particle breaks into the smaller globules at a much shorter distance than the sucrose particle. In addition, the shape of the disc is more bag like than in the sucrose trials due to the greater effect of the low pressure region present at the aft of the particle combined with the reduction in the particle viscosity. The smaller globules that result after primary

breakup of the initial particle are consequently less viscous, and therefore, the secondary breakup is rapid and is dominated by the dissolving action of the ambient fluid.

As the viscosity ratio is reduced, the shape of the particle becomes even more non-spherical as indicated in Figure 8 showing a sucrose particle at a VR of 144 and a Re of 15, and in Figure 9, showing an oil particle with a VR of 157 and a Re of 42. Due to the decreased initial internal viscous force, both particles are influenced by the low pressure region aft of the particle and the resulting shape is that of a bag. However, the sucrose particles developed a much more distinct bag shape, and it was an identifiable breakup mode for a VR of 20 to 200. As the sucrose bag traveled through the test section, the periphery would thin out until the viscous forces were no longer able to maintain a coherent shape, thus causing the periphery to collapse inwards under the pressure differential between the leading and trailing edges. The mass that had constituted the thin periphery would then tear away from the main particle and form a cloud of slow moving mass that dissolved quickly. The center of the bag, which contained a concentrated amount of mass, would then exhibit the viscous globule behavior described earlier.

For the oil particles, the bag behavior existed only in the vicinity of a VR of 150, and as indicated in Figure 5, the oil case tends to breakup at a distance less than that of the sucrose data. Furthermore, due to differences in the effect of solubility between the two different particle fluids (oil and sucrose), the thinning action of the bag development caused the oil particle to be dissolved faster than in the sucrose particle. Therefore, upon breakup of the primary oil particle, the entire amount of mass exhibited a puff-like behavior, and it quickly dissolved. The bag mode observed in the oil/pentane experiments did not constitute a primary breakup regime as it did in the sucrose experiments due to the bag's shorter life, reduced incident of occurrence, and greater ambiguity. Rather, all of the globule and bag modes were grouped into the general viscous regime denoted in Figure 4.

At viscosity ratios of approximately 30, the bag behavior is still the primary breakup mode for the sucrose particles, as seen in Figure 10. However, at the same viscosity ratio, the oil particles exhibit a vortex ring behavior as illustrated in Figure 11. Full vortex ring behavior was observed for the oil/pentane trials at a VR of 120 and a Re of 95. The vortex ring structure is similar to that observed Peck and Sigurdson [5] for a water droplet falling into water. The vortex ring is initiated by the impact with the free surface, which creates a discontinuous velocity gradient within the particle as a result of the shear forces acting on the particle fluid that is in contact with the ambient fluid. The particle fluid which is not in contact with the ambient fluid continues downward at a greater velocity until it comes in contact with the ambient fluid. The result is a recirculation vortex that is stabilized by the pressure differential between the leading and trailing edges of the azimuthal ring. At the higher viscosity ratios discussed previously, the internal viscous forces of the particle were sufficient to dampen out the differences in inertial forces between the mass in contact with the ambient fluid and the internal mass of the particle. Therefore, the vorticity that would result was diminished.

The oil particles exhibit vortex behavior at higher viscosity ratios and lower Re than the sucrose particles due to solubility. As stated earlier, the solubility effect reduces the viscosity of the particle fluid, and thus

diminishes its ability to dampen out the vorticity created by the inertial and pressure differences within the particle. Even for a viscosity ratio as low as VR=14, as shown in Figure 12, the sucrose particles have still not attained the vortex structure. At that point, the primary mode is neither vortex nor bag dominated. This was therefore termed as the transition regime. In Figure 12c gross vortex motion is observable, however, the particle mass quickly dissolves and the motion never becomes sustained. Figure 13 illustrates the stability of the vortex ring for the oil particles at a VR of 14. It should be noted that the Reynolds number is 320 for Figure 13, whereas for the sucrose particle of Figure 12, Re was only 170. At similar viscosity ratios, the Reynolds number is much greater for the oil/pentane experiments than the sucrose/water case due to fact that while the initial velocities and diameters of the two particle fluids are similar, the oil viscosity is 80% less than that of the sucrose.

At a VR of 7.5 and a Re of nearly 300, the sucrose particles develop a stable vortex ring. The VR at which the vortex behavior is dominant for the sucrose/water combination is over 90% smaller than the value for the oil/pentane combination. Similarly, the Re number that begins the vortex regime for the sucrose particle is 60% greater than that of the oil case. This is a clear demonstration of the major role that solubility plays in characterizing the deformation and breakup modes of a zero surface tension particle. Furthermore, the nature of the vortex mode is also different between the two fluids. As shown in Figure 14, the presence of mass nodes within the vortex structure is distinct, and it was observed to be the dominant factor governing the depth at which the vortex convected through the test section before breaking up. The mass nodes were not as distinct in the oil vortex rings, and while the dynamics responsible for the nodes has not yet been sufficiently explained, it is likely that the presence of solubility will determine to what degree they are present.

Piezoelectric LOX Droplet Generator

The droplet generator readily produced a stream of spherical LOX droplets, as typified in Figure 15. However, the uniformity of size within the stream, and the spacing between droplets exhibited instabilities. Figure 16 is a plot of the length and the Sauter mean of the LOX droplet stream for a wide range of frequencies. The average sample size for the different frequencies was 5200 drops. The striking feature of the data is that the length mean is relatively constant over the span of frequencies for large droplet sample sizes. This is a reasonable result and it corresponds to the Rayleigh breakup theory which indicates that for a single value of the jet diameter, there is an optimum value of the wavelength of the perturbations required to induce a stable breakup. The wavelength is a ratio of the jet velocity and the oscillation frequency, therefore, a change in one parameter must be accompanied by the corresponding change in the other to maintain the optimum wavelength. Thus, increasing the frequency of oscillation and the mass flow rate of the oxygen for a constant orifice diameter will result in a longer jet and an increased droplet velocity. However, to alter the diameter of the droplets markedly, the orifice diameter should be altered.

The variation in the Sauter mean reflects the variance in droplet diameters, whereas the length mean damps out influence of the larger droplets over a large sample size. The fluctuation in droplet sizes can be attributed to three sources. First, to achieve the flow rate of oxygen required for the optimum perturbation wavelength, a more precise control mechanism than the current flow control valve is needed. Second, it has been suggested by Dr. Talley's group at Phillips laboratory that the location at which the condensation of the GOX occurs may fluctuate enough to incur instabilities in the droplet stream. This effect can be reduced by increasing the length of the oxygen line within the heat exchanger. Finally, the orifice used contained many blunt surfaces and therefore, the optimum disturbances induced by the piezoelectric device may be damped out or altered. To reduce this effect, an orifice with smoother contours should be implemented. The next step in this study would be to characterize the droplet velocities and diameters as a function of the orifice diameter and the piezo oscillation frequency.

5.   Conclusions

As consequences of supercritical conditions, the surface tension vanishes and solubility between the gas and liquid phase increases. In simulation of this phenomenon, two miscible fluids were utilized to create zero-surface tension particles (droplets). In the absence of surface tension, the primary forces present during the deformation and breakup of a particle are the internal viscous forces and the inertial forces. The breakup regimes of sucrose particles falling into water and oil particles falling into pentane are given as a function of the VR and the Re. The primary regions observed are the viscous regime, transition, and vortex behavior. However, the point at which the various regimes occurred varied between the sets of fluids as a result of the solubility, which tends to alter the viscosity of the particle fluid. Therefore, to characterize the breakup regimes of a zero surface tension droplet, and presumably a droplet undergoing transition in a supercritical environment, the solubility rate must be included with the viscosity ratio and the Reynolds number.

For the piezoelectric LOX droplet generator, the length mean diameter of a large sample size of droplets remained constant at approximately 250 μm over a wide range of piezo oscillation frequencies for the six μm orifice used. This corresponds to the Rayleigh breakup mode which indicates that the diameter of the droplets is a function of the jet diameter. Thus, to alter the jet diameter, the orifice of the generator must be changed. However, the stream of droplets was not entirely uniform in size, and these variations are indicated in the fluctuation of the Sauter mean diameter. The instabilities present resulted from the inability to precisely control the mass flow rate of the oxygen, the variation in the location at which GOX condensation occurs, and blunt corners of the orifice. As these issues are addressed, the generator will become an invaluable resource in supercritical droplet breakup study.

6. References

1. Talley, D. G., R. D. Woodward, and J. C. Pan, "Supercritical Droplet Behavior," AFOSR Task 2308AP (1995).

2. Probstein, R. F., Physicochemical Hydrodynamics," Butterworths, 1989.

3. Lefebvre, A. H., *Atomization and Sprays*, Hemisphere Publishing Corporation, N.Y., 1989.

4. Woodward, R. D., D.G. Talley, T.J. Anderson, and M. Winter, "Shadowgraphy of Transcritical Cryogenic Fluids," presented at the 6th Annual Symposium of the Pennsylvania State University Propulsion Engineering Research Center, NASA LeRC, Cleveland OH, (1994).

5. Peck, B., and L. Sigurdson, "The Vortex Ring Velocity Resulting From an Impacting Water Drop," *Experiments in Fluids*, vol. 18, pp. 351-357, 1995.

Macintosh Power PC
7100/80 With Frame
Grabber Board

Droplet Generator Assembly

Strobe Light

Stanford Optics
CCD Camera
With Image
Intensifier

Test Section

(a)

Droplet Generator
Containing
Particle
Fluid

Calibrated Rail For
Easy
Adjustment of Droplet
Generator Height

h

Free Surface

Pexiglass Test
Section

Ambient
Fluid

(b)
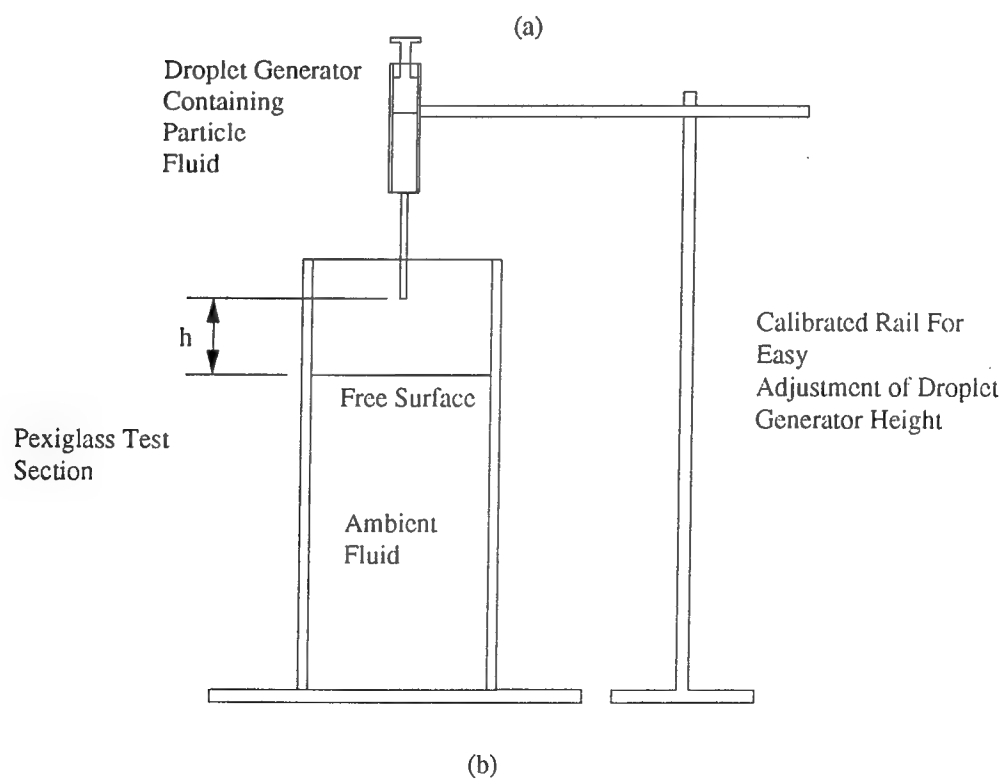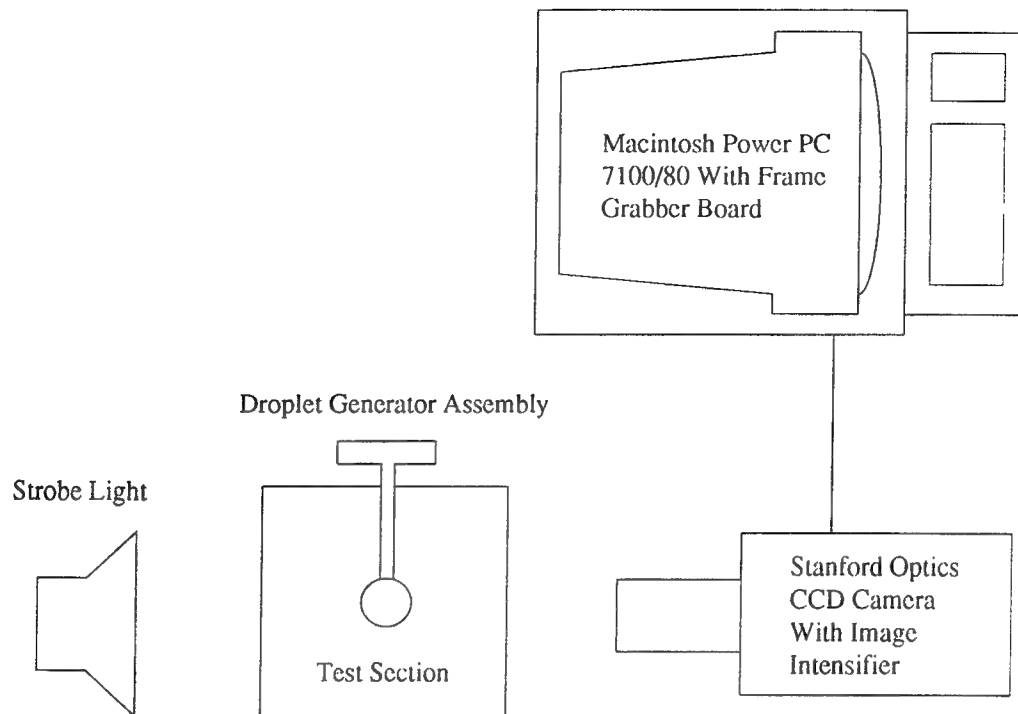
Figure 1. (a) Schematic of experimental setup (b) Schematic of droplet generator and test section
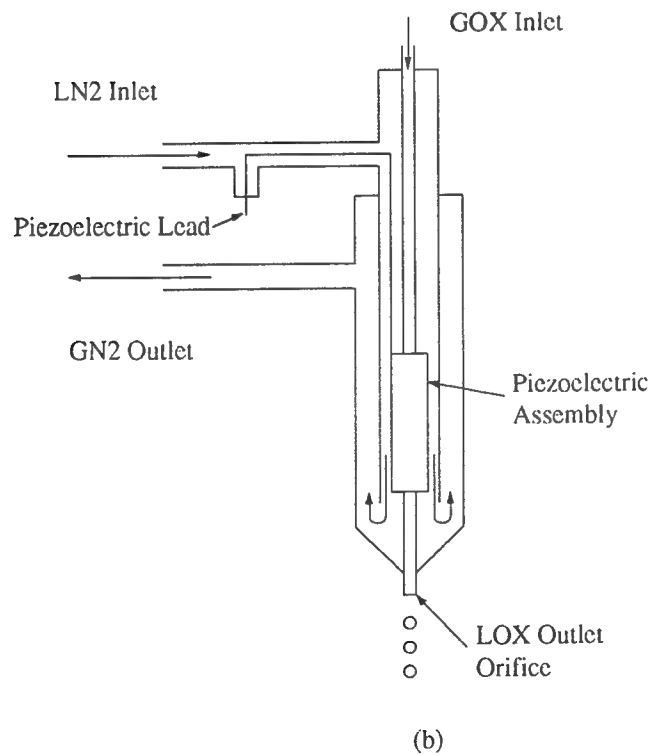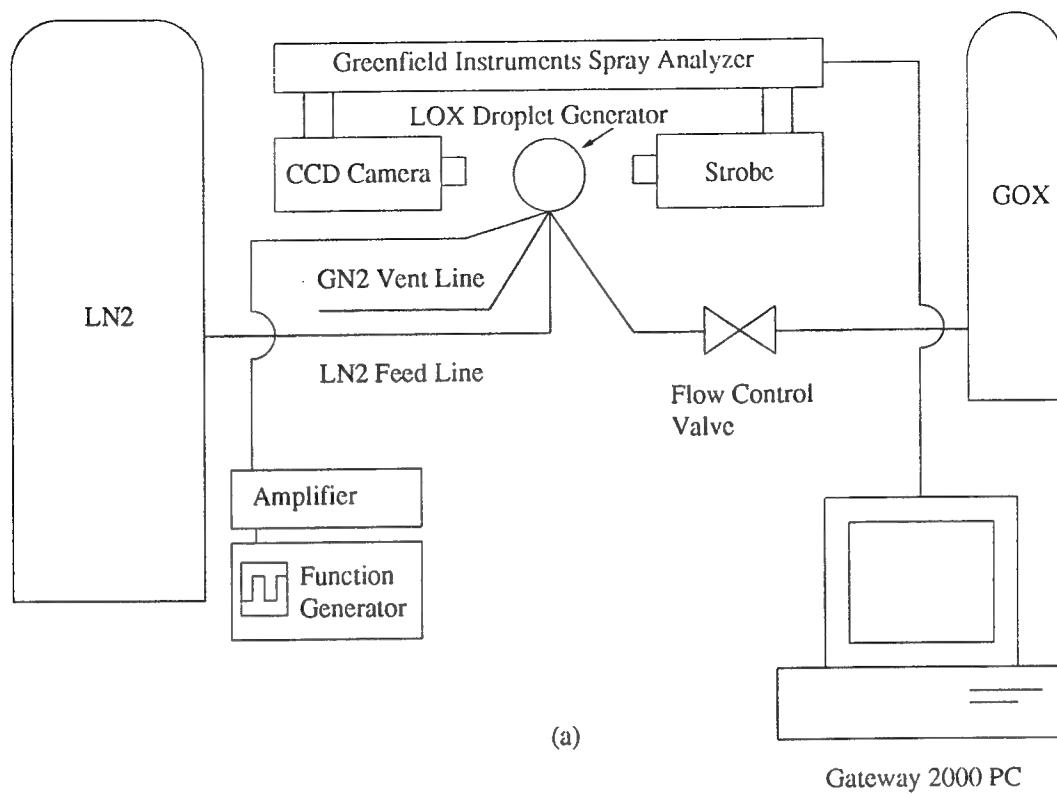
Figure 2. (a) Schematic of LOX droplet generator experimental setup (b) Schematic of LOX droplet generator
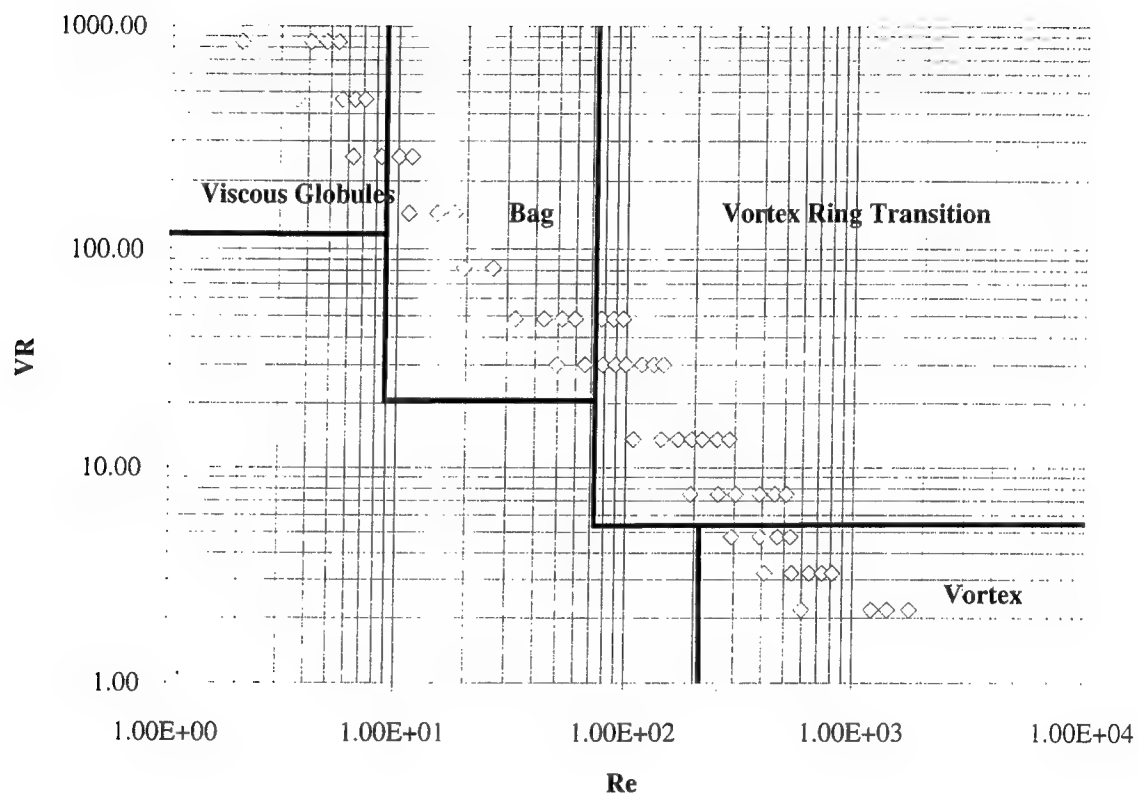
Figure 3. Primary breakup modes of a zero surface tension sucrose particle as a function of VR and Re
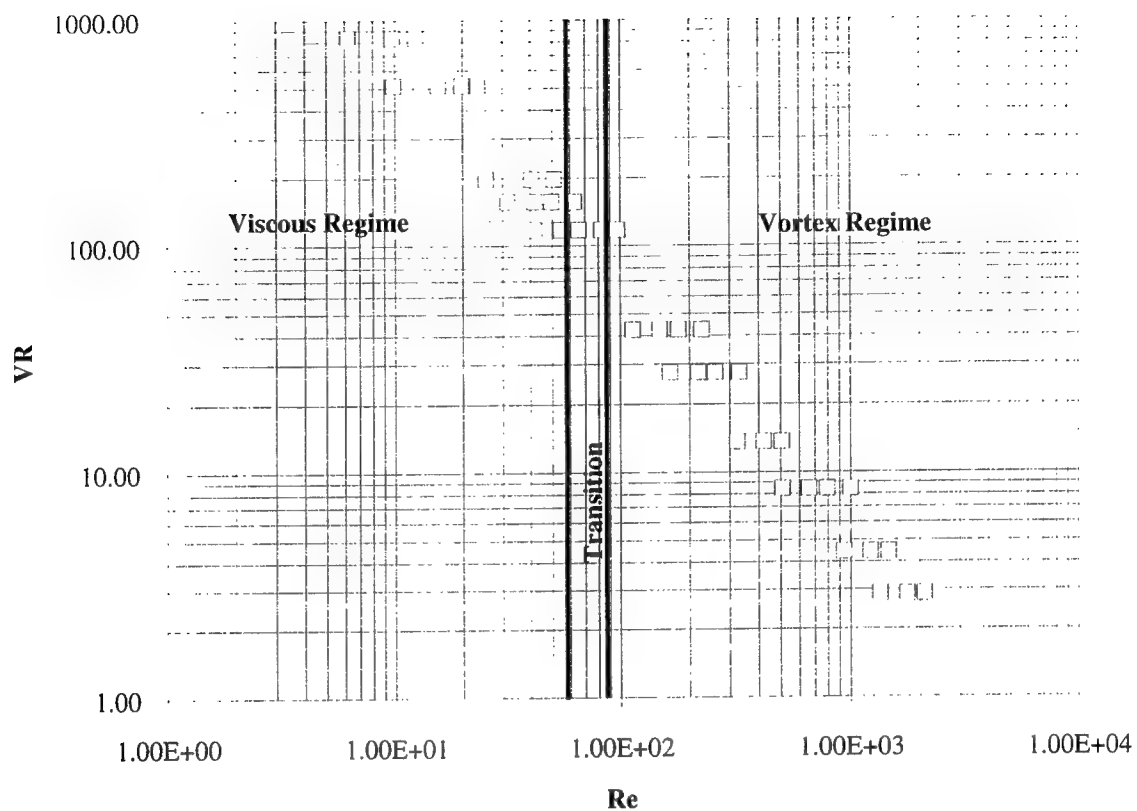
Figure 4. Primary breakup modes for a zero surface tension oil particle as a function of VR and Re
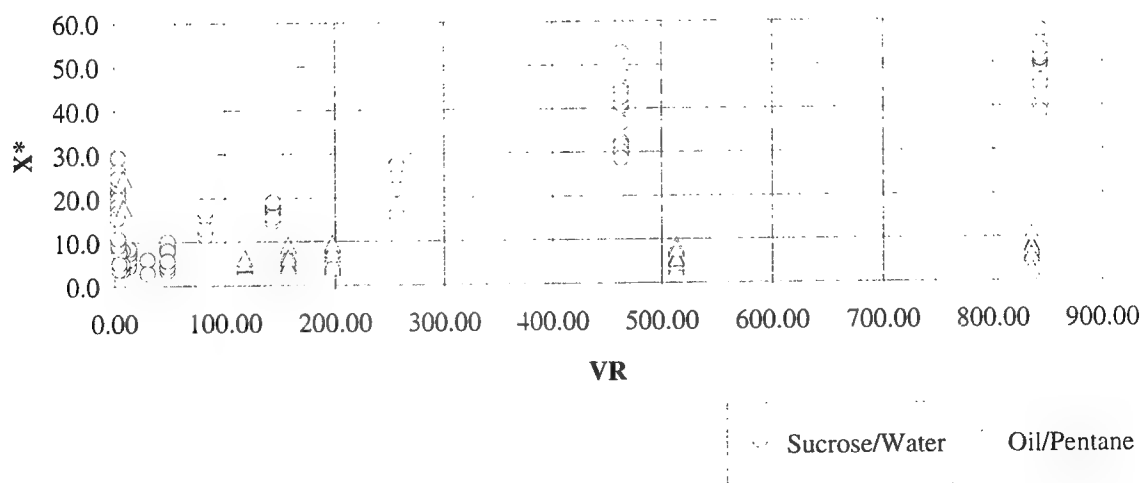
Figure 5. Nondimensional distance to primary breakup, X*, as a function of VR

(a)

(b)

(c)
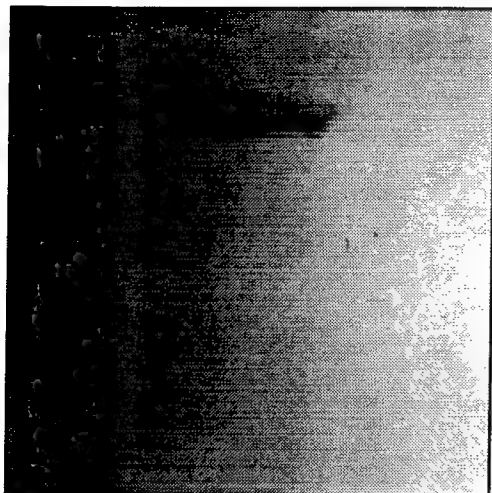
(a)
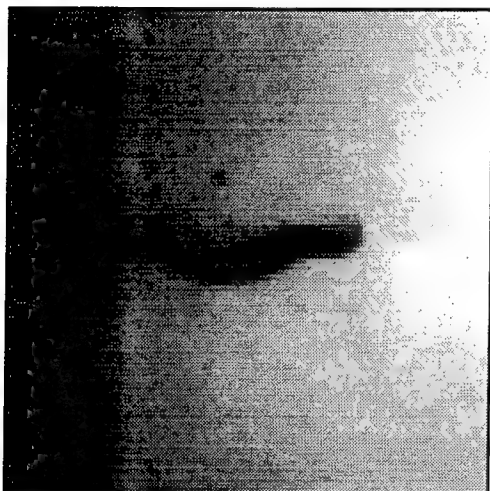
(b)
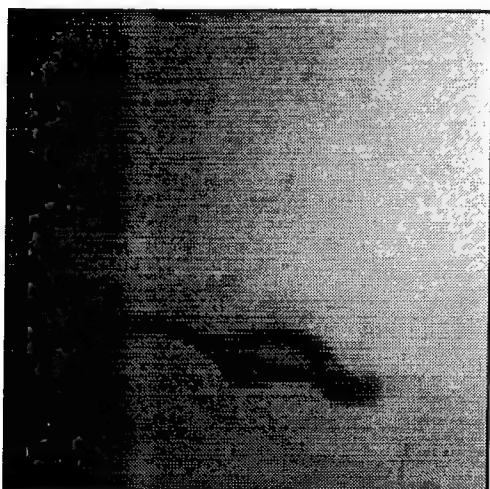
(c)

Figure 6. Deformation and breakup of a zero surface tension sucrose particle falling freely in water with initial conditions of VR=843 and Re=5
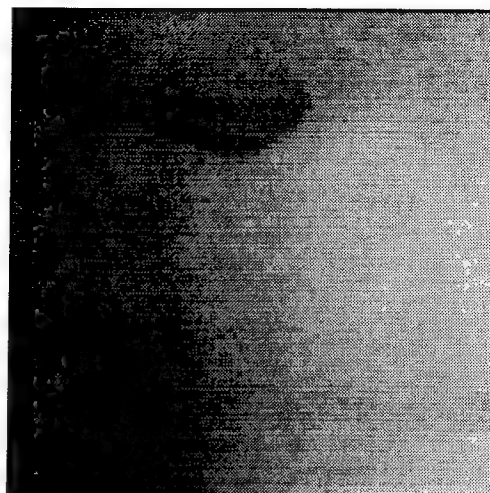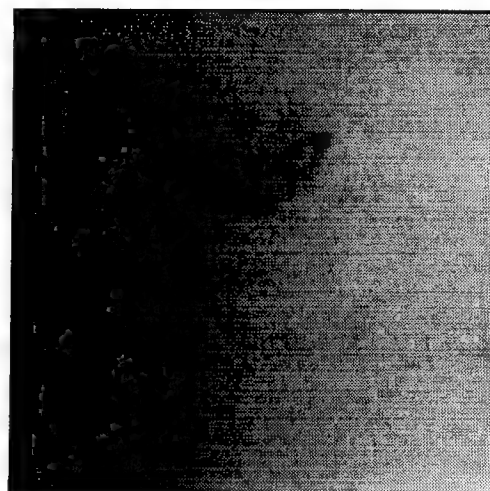
Figure 7. Deformation and breakup of a zero surface tension oil particle falling in pentane with initial conditions VR=835 and Re=6
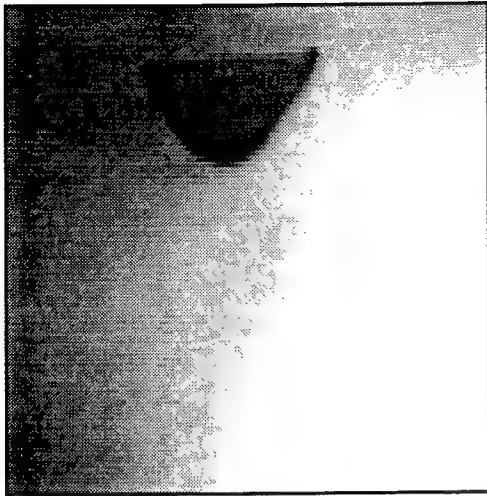
(a)

(b)

(c)

Figure 8. Deformation and breakup of a zero surface tension sucrose particle falling in water with initial conditions of VR=144 and Re=15
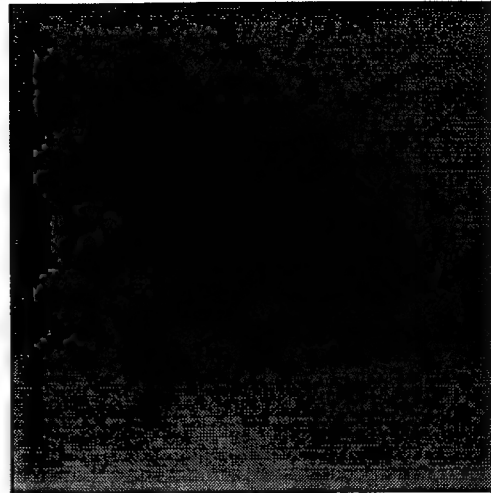
(a)
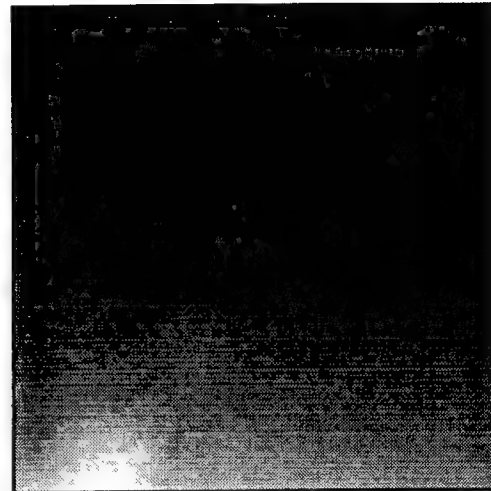
(b)

(c)

Figure 9. Deformation and breakup of a zero surface tension oil particle falling in pentane with initial conditions of VR=157 and Re=42
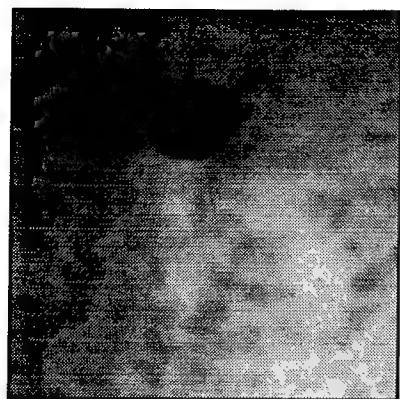
(a)
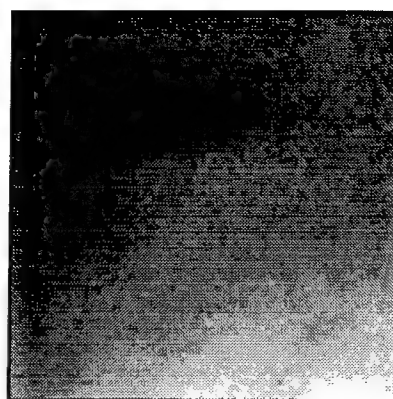
(a)
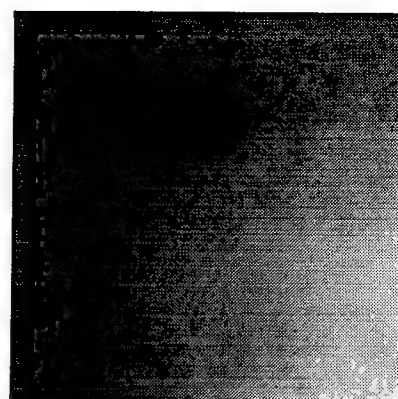
(a)
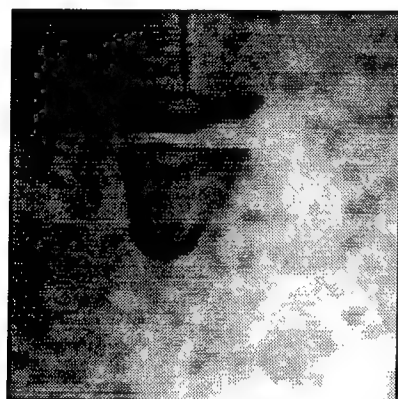
(b)
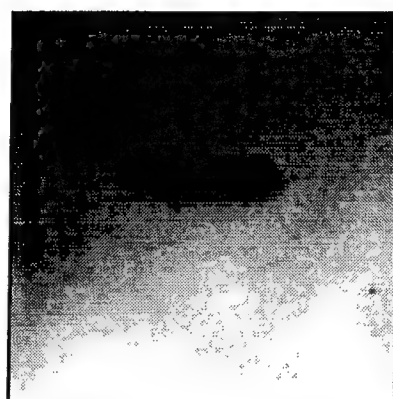
(b)
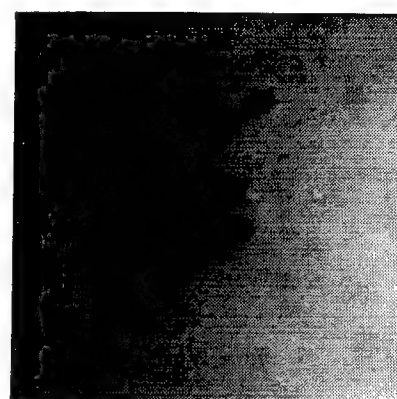
(b)

(c)
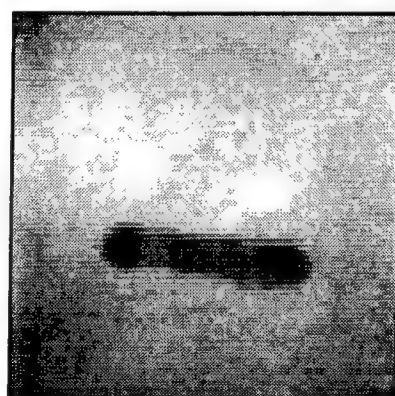
(c)
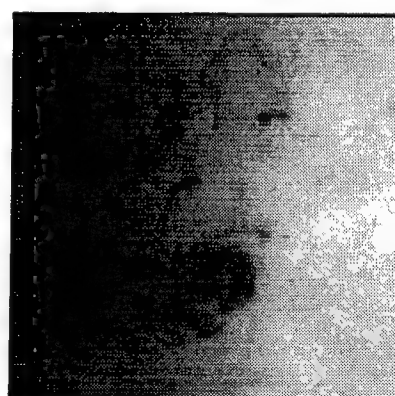
(c)

Figure 10. Deformation and breakup of a zero surface tension sucrose particle falling through water with initial conditions of VR=30 and Re=66

Figure 11. Deformation and breakup of a zero surface tension oil particle falling through pentane with initial conditions of VR=28 and Re=260

Figure 12. Deformation and breakup of a zero surface tension sucrose particle falling through water with initial conditions of VR=14 and Re=170
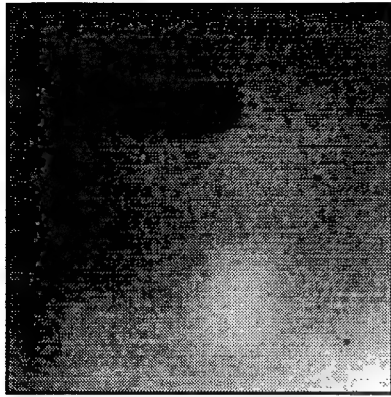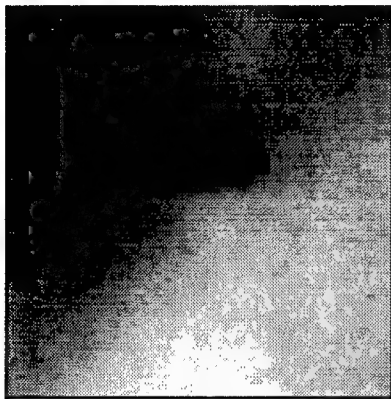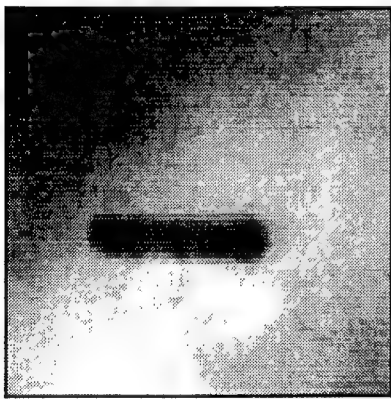
Figure 13. Deformation and breakup of a zero surface tension oil particle falling through pentane with initial conditions of VR=14 and Re=321
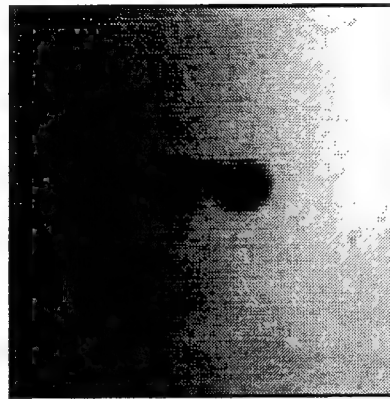
Figure 14. Deformation and breakup of a zero surface tension sucrose particle falling through water with initial conditions of VR=3.2 and Re=734
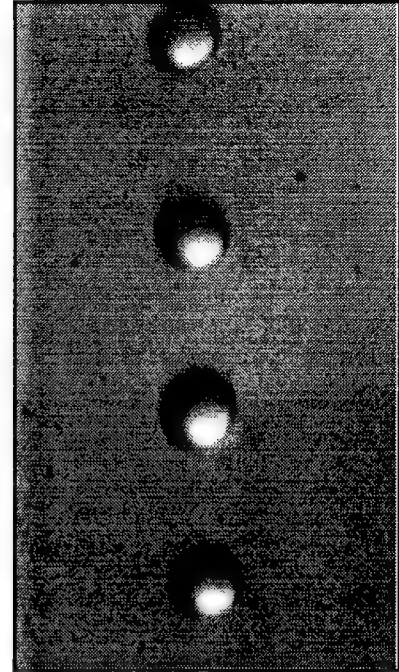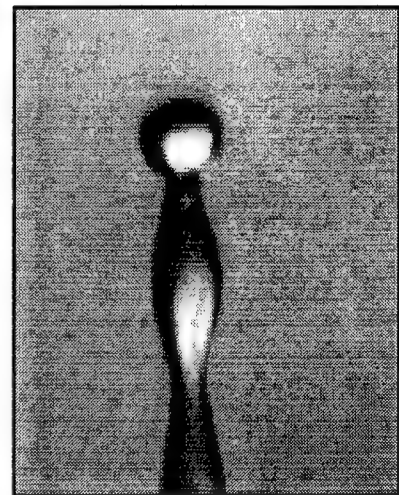
Figure 15. (a) Stable LOX droplet stream (b) LOX jet from droplet generator illustrating the Rayleigh breakup mode induced by the piezo oscillations
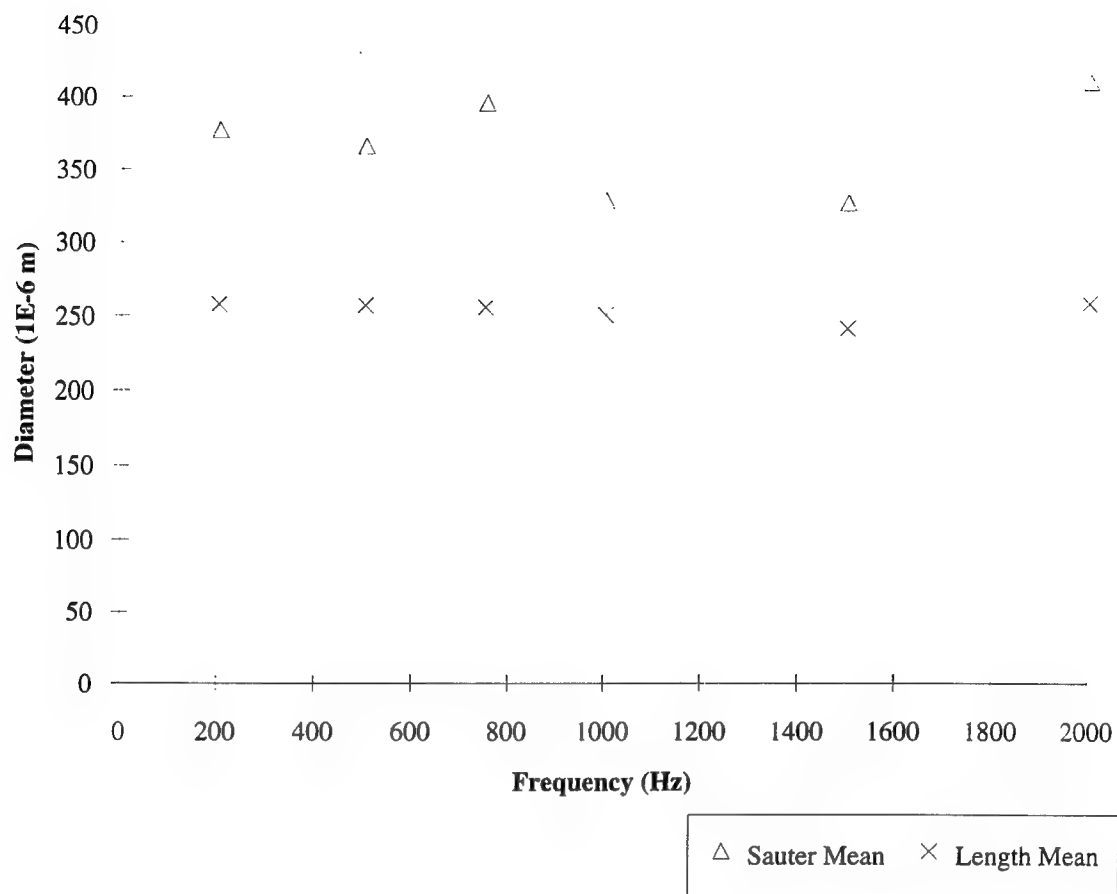
Figure 16. Length and Sauter mean diameters as a function of piezo oscillation frequency for a the LOX droplet generator with a 6 micron orifice

A C++ TEST BED FOR GRAPH PARTITIONING AND MAPPING FOR PARALLEL PROGRAMMING

Mark Purtill
Assistant Professor
Department of Mathematics

Texas A&M University–Kingsville
Campus Box 172
Kingsville TX 78363

A C++ TEST BED FOR GRAPH PARTITIONING AND MAPPING FOR PARALLEL PROGRAMMING

Mark Purtill
Assistant Professor
Department of Mathematics
Texas A&M University–Kingsville

Abstract

Many problems in computational physics can be viewed as computations carried out on graphs (networks); one such problem arises in particle-in-cell (PIC) plasma simulations. As many multiprocessors (massively parallel computers) can also be viewed as graphs, the problem of partitioning such problems to run efficiently on a multiprocessor is a graph theoretic problem. It can be viewed as choosing a graph map between the two graphs minimizing a function.

Unfortunately, this problem is NP-complete, which means it is very unlikely that we can find an efficient, exact algorithm. Therefore, we must rely on heuristic methods. Many such methods have been proposed for this and related problems.

To investigate which of these algorithms are best suited for the graphs arising from PIC simulations, we have developed a test bed program which runs on both networks of workstations and on multiprocessors. The test bed is written in C++, using modern programming methods such as object-oriented design and literate programming.

A C++ TEST BED FOR GRAPH PARTITIONING AND MAPPING FOR PARALLEL PROGRAMMING

Mark Purtill

1 The Problem

1.1 Graph Theory Background

A *graph* G consists of a set of objects, called *vertices* (singular *vertex*), denoted $V(G)$ and a set of unordered pairs of vertices, called *edges*, denoted $E(G)$. (For our purposes, both of these sets must be finite.) The two vertices in an edge are called its *ends*, and the edge is said to *connect* its ends. We do not allow an edge to connect a vertex to itself, nor for more than one edge to connect a given pair of vertices (that is, our graphs are "simple").

A graph may be represented by a drawing (picture), such as figure 1; however, the geometric information (such as coordinates of the dot representing a vertex) are *not* part of the graph; figure 2 represents the same graph G as figure 1: the vertex set is $V(G) = \{a, b, c, d\}$ and the edge set $E(G) = \{\{a, b\}, \{a, c\}, \{b, c\}, \{b, d\}, \{c, d\}\}$.

If every pair of vertices is an edge (that is, all pairs of vertices are connected), then the graph is called "complete". Figure 3 is a complete graph on five vertices.

Note that in some cases (such as figure 3), it is necessary for the edges of the drawing to cross; this has no effect on the graph itself. Graphs that can be drawn without crossings are called planar graphs; these have various nice properties (see *e.g.* [9]). We do not assume our graphs are planar, since many interesting graphs are not planar.

A sequence of vertices $(v_1, v_2, v_3, \ldots, v_n)$ such that each pair of vertices v_i, v_{i+1} is called a *path*. For instance, in the graph in figure 1, (a, c, b) is a path, while (a, d, b) is not. (Note that we can also specify a
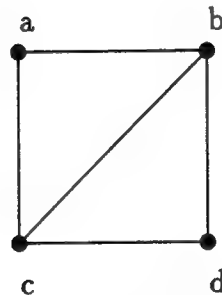

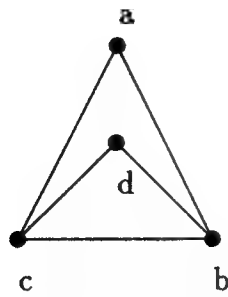Figure 1: A drawing of a graph; the disks represent vertices, and the lines represent edges.

Figure 2: A different drawing of the same graph as in figure 1.
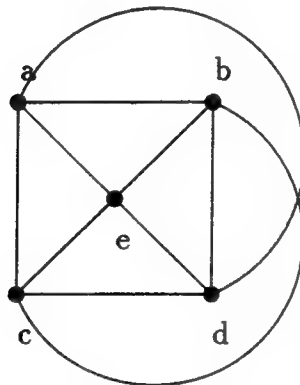

Figure 3: A drawing of a graph which cannot be drawn without edges crossing: the complete graph on five vertices.
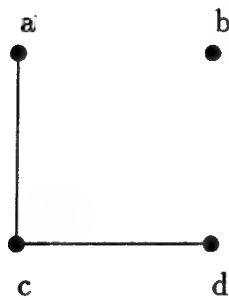

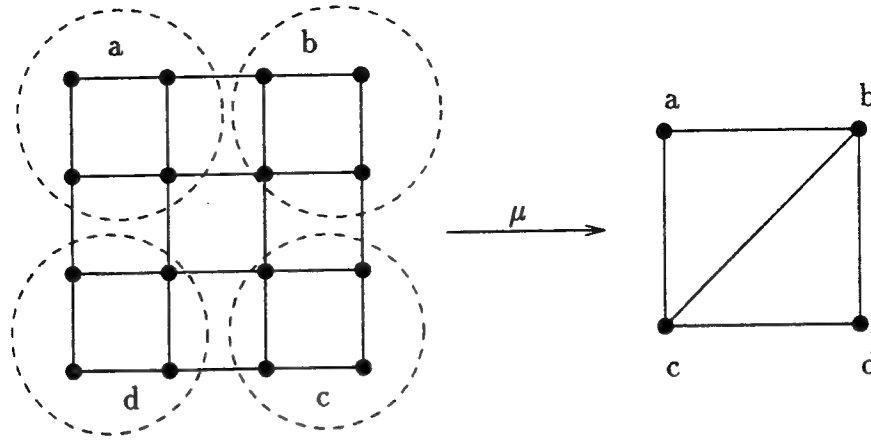Figure 4: This graph is disconnected because (*e.g.*) there is no path from vertex *a* to vertex *b*.

Figure 5: A graph map; each of the dashed circles in the graph on the left hold the vertices in the preimage of the vertex in the graph on the right with the same label.

path by listing the edges in that path rather than the vertices.) The *length* of a path is the number of *edges* in it.

If there is a path between any pair of vertices in the graph, then the graph is called *connected*. The graphs in figures 1, 2 and 3 are connected; that in figure 4 is not connected. All graphs considered in this paper will be connected. The *distance* between two vertices in a graph is the length of the shortest path connecting them; the *diameter* of the graph is the longest distance between two vertices. For instance, the graph in figure 1 has diameter 2.

A *graph map* μ from a graph G to a graph H, denoted $\mu: G \to H$ is a function μ from the vertex set of G, $V(G)$, to the vertex set of H, $V(H)$, so that for every edge $\{v, w\}$ of G, either $\mu(v) = \mu(w)$ or $\{\mu(v), \mu(w)\}$ is an edge of H; that is:

$$\text{if } \{v, w\} \in V(G), \text{ then either } \{\mu(v), \mu(w)\} \in V(H) \text{ or } \mu(v) = \mu(w).$$

(Aside for graph theorists: if we assume each vertex of each graph is equipped with a self-loop, this coincides with the usual definition of *e.g.* [9].)

For each vertex v of H, we denote by $\mu^{-1}(v)$ the set of all vertices of G which map under μ to the vertex v (the *preimage* of v); that is $\mu^{-1}(v) := \{ w \in V(G) \,|\, \mu(w) = v \}$. A sample graph map is shown in figure 5.

The *valance* or *degree* of a vertex is the number of edges of which it is an endpoint, or, equivalently, the number of vertices which are connected to it by an edge. In figure 1, vertices a and d have valance 2, which b and d have valance 3. If all of the vertices of a graph have the same valance, the graph is called *regular*; for instance, figure 3 shows a regular graph (of valance 4).

For a simple application of the graph concept, consider a multiprocessor; that is, computer consisting of a number of different processors, each with its own memory, which can communicate with each other *via*
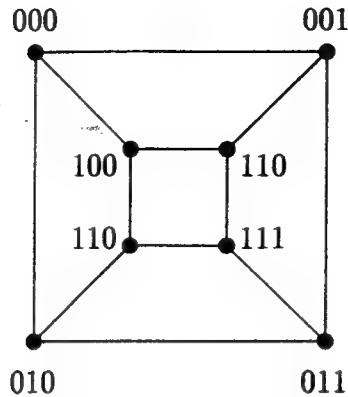
Figure 6: A three-dimensional hypercube graph.

communications links. Assume each link connects exactly two processors.

This arrangement of processors can be expressed as a graph; the vertex set of the graph is the set of processors, while the edge set is all the pairs of processors which are directly connected. (So the set of edges is the same as the set of links.) When we think of a graph as representing a multiprocessor, we will refer to the vertices (edges) as processors (links).

A path $(p_1, p_2, \ldots, p_n)$ in this graph is a way that a message could be passed from processor p_1 to processor p_n through processors p_2 through p_{n-1} *via* direct links, and the graph is connected precisely when each processor can communicate with each other processor (directly or indirectly).

For example, a common network connection is the "n-dimensional hypercube", which consists of 2^n processors, labeled 0 to 2^{n-1}. Two processors are connected if the binary representation of their labels differ in exactly one bit. See, for example, figure 6. For instance, the computer I did some work on this problem in 1994 (an Intel iPSC/860 called neutrino—see [21]) had the topology of the four-dimensional hypercube.

Often it is handy to be able to associate a number with each vertex or each edge of the graph; we call such numbers *weights*. Vertex weights are numbers associated with each vertex of the graph, while edge weights are numbers associated with each edge. For instance, we might have vertex weights giving the relative speed of each processor, and edge weights giving the speed or bandwidth of each communications link.

Formally, if we say a graph G has "vertex weights w", we mean w is a function from the vertex set $V(G)$ of G to the real numbers, and write $w(v)$ for the weight of vertex v. Similarly, if we say G has "edge weights u", we mean u is a function from the edge set $E(G)$ of G to the real numbers. In this paper, we will only need vertex and edge weights which have non-negative values.

1.2 Motivation from Plasma Physics Simulation

Many problems involving numeric solving of partial differential equations can be cast as graph computation. Here we describe one such problem, the particle-in-cell method for numerical computation of medium density
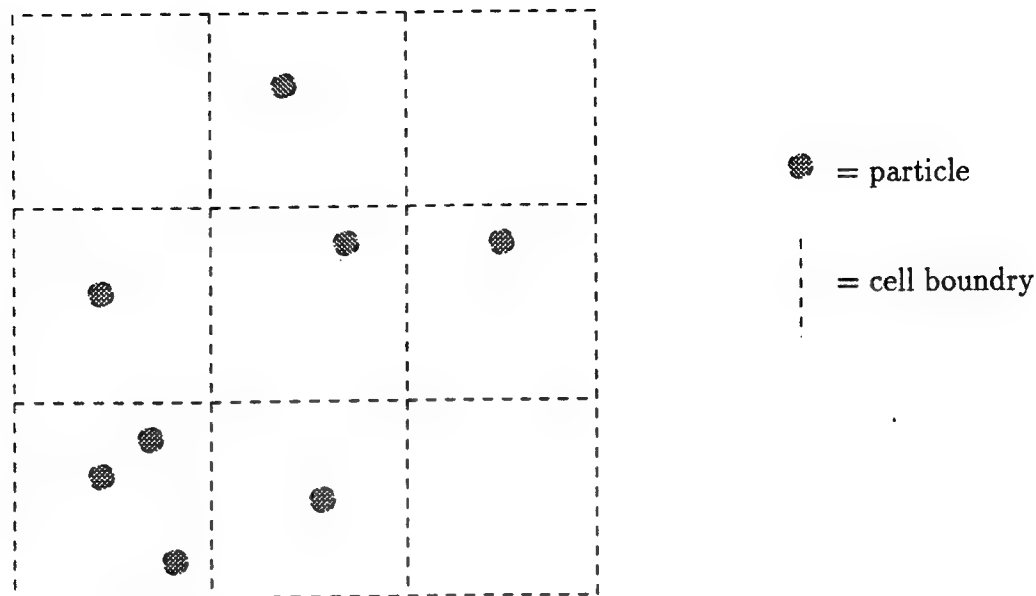
Figure 7: Particle-in-cell plasma simulation.

plasma (PIC plasma simulation).

In a plasma, there are magnetic and electrical fields interacting with charged particles moving in space; the interaction is governed by Maxwell's equations (for the fields) and the Lorentz equations (for the motions of the particles). In the particle-in-cell model, space is divided into cubes (cells) by a grid. The fields are kept track of on the grid points (or, in more sophisticated codes, on faces or edges of the cubes, with different field components having different locations). The particles, on the other hand, are allowed to move freely; they are kept track of by which cell they are in, then the coordinates of the particle within the cell. For more details, see (for instance) Birdsall and Langdon [2].

We can construct a graph to represent this computation. The vertex set of the graph is the set of cells of the spatial decomposition, while the set of edges is the pairs of cells which must exchange information in order for the computation to proceed. This information exchange occurs two ways in this problem:

- particles sometimes will move between adjacent cells; and
- adjacent cells must exchange field information in order to "step" Maxwell's equations.

Depending on the exact details, and whether it is necessary to keep track of both kinds of communication, the exact details of the edges may vary; see figure 8 for a typical example. (In this case, we assume that particle movement between diagonally connected cells is sufficiently rare that transferring the particle *via* another cell is acceptable. Another option would be to use edge weights.)

The computation graphs we get from PIC codes will have a nice structure; for instance, "most" cells (all those not on the boundary) will have the same valance; thus, the graph is "close" to being regular. In the case of figure 8, for instance, the center cell is "typical" (in the sense that a finer subdivision of space would give
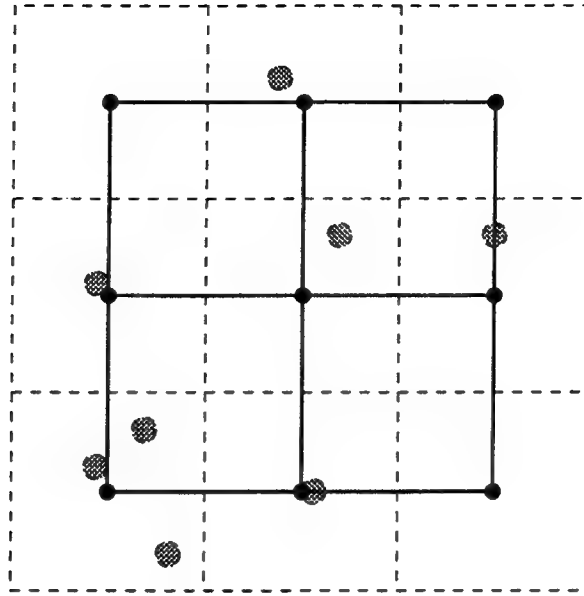
Figure 8: A two-dimensional subdivision of space (dashed), together with a corresponding graph. Note that the even distribution of space does not result in an even balance of particles.

a large number of cells which look like that cell); its valance is 4. (It is not clear whether this near-regularity would be useful.)

Because of the movement of particles, PIC simulations have an additional feature compared with some other problems: the amount of work to be done at the vertices representing the cells will change over time as particles move in the system. The amount of computation needed will be (in general) will be unpredictable. (At a given time step, amount of computation needed could be given by vertex weights.)

Now, suppose we wish to run a PIC plasma simulation on a multiprocessor. For maximum efficiency, we would like to assign cells to different processors in such a way that adjacent cells map to either the same processor or processors connected by a communications link, in such a way that processors have (as close as possible to) equivalent load and not too much communications is necessary.

This can be reformulated in the language of graph theory: the assignment of cells to processors is exactly a graph map μ from the PIC plasma simulation's graph G to the multiprocessors graph M; this is the statement that adjacent cells map to the same or connected processors. If the processors and links of the multiprocessor are all the same, and the vertices and edges of the PIC plasma simulation graph have equal load on them, then the statement that the processors are equally loaded reduces to requiring that the sizes of the $\mu^{-1}(v)$ be the same, and the requirement that there be a small amount of communications becomes the requirement that there be a small number of edges connecting the $\mu^{-1}(v)$s. In case of differing speeds of processors and so on, vertex and edge weights can be used to capture the differences.

The problem we are mainly concerned with is a generalization of the problem just discussed; we assume

34-8

that there is already a map μ, which is not necessarily optimal, and we wish to find a new map ν which is load balanced, has low communication, and is close to μ. If we think of the map μ as an assignment of vertices of the PIC graph to processors of a multiprocessor, then this last desire says that not too many vertices must be moved between processors to change to the new assignments. The relative weightings of the three desires would depend on a number of factors; for instance, if the problem is near completion, we would not be willing to spend a lot of time to speed up the last few steps, and thus the last desire would become more important; at the begining of a long run, it would be less important or even ignored entirely.

1.3 Mathematical Formulation

We can formulate this "Graph Remapping" problem in mathematical terms. Using techniques of [8], we can reduce it even the "mapping" problem to an "NP-complete" problem; this means it is extemely unlikely that there is an algorithm which solves the mapping (or remapping) problem optimally in time polynomial in the size of the input.

Consider two graphs, G and M.

The graph G, representing the problem to be solved, has vertex weights w and edge weights c: for a vertex $v \in \mathrm{V}(G)$, $w(v)$ is the amount of work needed to be done at v during one step of the computation; for an edge $e \in \mathrm{E}(G)$, $c(e)$ is the amount of communications needed between the two ends of the edge. Borrowing from graph embedding, we will call G the "guest graph".

For the graph M, representing the multiprocessor the problem will be solved on, we call the vertices *processors* and the edges *links*. The graph M has vertex weights s and edge weights b: for a processor $p \in \mathrm{V}(M)$, $s(p)$ is the speed of processor p; for a link $l \in \mathrm{E}(M)$, $b(l)$ is the bandwidth of the link l. Again, borrowing from graph embedding, we will call M the "host graph"[1].

Given a graph map $\mu\colon G \to M$, the problem is to find a new graph map $\nu\colon G \to M$ minimizing some increasing function of

- $move_v = \sum\limits_{v \in \mathrm{V}(G)} d_M(\mu(v), \nu(v))$ for each $v \in \mathrm{V}(G)$;

- $work_p = \dfrac{1}{s(p)} \sum\limits_{v \in \nu^{-1}(p)} w(v)$ for each $p \in \mathrm{V}(M)$; and

- $comm_e = \dfrac{1}{b(l)} \sum\limits_{e \in \nu^{-1}(l)} c(e)$ for each $l \in \mathrm{E}(M)$.

For instance, the function f given by

$$f(\nu) = \max(\alpha\, move_v) + \max(\{\, \beta work_p \,|p \in \mathrm{V}(M)\,\} \cup \{\, \gamma comm_e \,|e \in \mathrm{E}(M)\,\}),$$

---

[1] This is also appropriate because M represents a multiprocessor "host".

34-9

with suitable choices of the constants α, β, and γ, could measure the total time needed to load balance (first term) and run a step (second term) on a multiprocessor which can overlap commutation and communications.[2]

We call the general problem "Graph Remapping"; we assume that the map μ represents a solution to the same problem at a previous time step. If the function to minimize takes no account of the previous map μ, then this reduces to the problem of "Graph Mapping"; this could be used to initially load the multiprocessor; alternatively, depending on the multiprocessors I/O capabilities, we might wish to start with a map μ which maps all the vertices of G to a single processor of M.

Because of the motivation from parallel computing, we are most interested in solving this problem on the multiprocessor modeled by graph M. Since communications is only possible on edges of the graph, any algorithm which balances the load requires at least the diameter of M communications steps. (Recall that the diameter of a graph is the longest distance between two vertices, or, in this case, processors. If few communications steps than the diameter are taken, then two processors which are the diameter apart cannot communicate, and thus cannot average loads with on another.)

1.4   NP-completeness

This algorithm can be considered as a decision problem by giving a bound F on $f(\nu)$ and asking if any mapping ν exists for which $F \geq f(\nu)$. In this form, the theory of NP-completeness can be applied to the problem. In the format of the standard reference in that area, Garey and Johnson [8], the problem would be expressed as follows (with our sample f given above):

MINIMAL GRAPH REMAPPING

INSTANCE: Graphs G and M and a graph map $\mu \colon G \to M$. Four nonnegative numbers K, α, β, and γ.

QUESTION: Is there a graph map $\nu \colon G \to M$ so that

$$f(\nu) = \max(\alpha\,move_v) + \max(\{\,\beta work_p\ |p \in \mathrm{V}(M)\,\} \cup \{\,\gamma comm_e\ |e \in \mathrm{E}(M)\,\}) \leq K\ ?$$

A decision problem (one with a yes-no answer) is said to be in class P if there is an algorithm which solves any instance of the problem in time which can be bounded by a polynomial in the size of the instance. Such an algorithm is called "a polynomial time" algorithm.

A decision problem is said to be in class NP if for any instance for which the answer to the problem is "yes", there is a polynomial sized "certificate" which can, again in polynomial time, be used to demonstrate that the answer is "yes". Our problem is in this class because the map ν itself is the certificate.

Now, $P \subseteq NP$ since if the problem can be solved in polynomial time, we can use the problem itself as the certificate. It seems intuitively clear that NP could be much larger than P, but so far no one has been able to prove it (or to disprove it, that is, to show that $P = NP$).

---

[2]Notice that even this expression is an approximation to the exact time, since (for instance) some processors could overlap moving of vertices with computation.

However, there is a nicely developed theory of "NP-completeness". Karp, in [15], showed that there are some problems which are the hardest problems in the class NP, in the sense that if we could solve one of these problems in polynomial time, then *all* problems in NP could be solved in polynomial time. Such problems are called NP-complete; if any one of them is in P, then $P = NP$. As the latter is considered unlikely, this is good evidence that there is indeed no algorithm to solve any NP-complete problem in polynomial time.

Furthermore, to show that a decision problem is NP-complete, it suffices to show that if a polynomial time algorithm was found for that decision problem, it could be used to develop a polynomial time algorithm for some specific NP-complete problem. See [8] for details.

For our problem, this is not hard. We will reduce our problem to the following:

MINIMUM CUT INTO EQUAL PARTS

INSTANCE: Graph G, positive integer K.

QUESTION: Is there a partition of V(G) into parts V_1 and V_2 such that V_1 and V_2 are of equal size, and no more than K edges pass between them?

which can be easily seen to be NP-complete by reduction to [8]'s ND17.

We can solve this problem by remapping by letting $\alpha = 0$ (so we don't need an initial map μ), letting all weights be 1, and taking the host graph M to be the graph with two vertices connected by an edge. We will use the same value of K in both problems. Thus, the graph map ν is just a partition of V(G) into two parts, and $f(\nu) = \max(\{\beta work_p \,|\, p \in V(M)\} \cup \{\gamma comm_e \,|\, e \in E(M)\})$ where now $work_p$ is just the number of vertices in each part of the partition, and $comm_e$ is the number of edges crossing between the parts.

So if we let $\beta = 2K/\#V(G)$ where $\#V(G)$ is the number of vertices of G, then only partitions into equal parts will satisfy $f(\nu) \leq K$. Setting $\gamma = 1$ completes the proof: the remapping will be possible only if at most K edges cross the partition.

2   Previous Work and Algorithms

2.1   Work on Graph Partition

The *graph partition* problem can be viewed as a specification of the graph mapping problem, in which the host graph is assumed to be complete (each vertex is connected to every other vertex). Typically, the vertex set of the graph is divided onto two equal parts V_1 and V_2; if more parts are desired, recursion is employed.

This problem been studied in, for example, [1,13,18,20,23], and several good heuristic algorithms have been found. We survey a few of the most popular.

2.1.1   Kernigan-Lin

The Kernigan-Lin algorithm [18] is a local improvement algorithm: given a partition, it attempts to make it better. There is a data structure [7] which allows Kernigan-Lin to run in linear time.

```
kl( V₁, V₂) {
    best := (V₁, V₂);
    best_cost := cost( V₁, V₂);
    mark all vertices as unfrozen;
    while( there are unfrozen vertices in both V₁ and V₂) {
        transfer( V₁, V₂);
        transfer( V₂, V₁);
        if( cost(V₁,V₂) < best_cost) {
        best := (V₁, V₂);
        best_cost := cost(V₁,V₂);
    }    }
    return best;
}
transfer( X, Y) {
    Choose v ∈ X minimizing cost( V₁ \ v, V₂ ∪ v);
    X := X \ v;
    Y := Y ∪ v;
    mark v as frozen in Y;
}
```

Figure 9: One step of the Kernigan-Lin graph partition improvement algorithm.

The Kernigan-Lin algorithm works by improving a partition of a graph into two parts, say V_1 and V_2. (So $V(G) = V_1 \cup V_2$, and $V_1 \cap V_2 = \emptyset$.) The *cost* of such a partition is the number of edges between V_1 and V_2.

One step of the algorithm begins by saving the current partition as the best-so-far, along with its cost. We then loop over the following: a vertex is transfered from V_1 to V_2; then a vertex is transfered back from V_2 to V_1. The resulting partition is compared with the best-so-far, with the new partition being saved as best-so-far if it is strictly better than the old one.

The vertex to be transfered from V_1 to V_2 is the vertex v which minimizes the cost of the partition $(V_1 \setminus v, V_2 \cup v)$ (regardless of whether this partition is an improvement over the previous one). That is, we move the vertex giving the best (or least bad) partition if it is transfered to the other part. However, once a vertex has been transfered once, it is "frozen" and can not be transfered again. The vertex transfered the other way is chosen similarly.

This step is shown in pseudocode in figure 9; the step would be called repeatedly until the partition does not change. Not that although partitions better than the current partition are considered during the step, the returned partition is at least as good as the input partition.

An interesting variant of Kernigan-Lin is the Mob heuristic of [23]. Essentially, instead of transferring vertices one-by-one, the Mob heuristic transfers vertices in groups; the size of the groups are chosen beforehand according to a "Mob schedule". Since it would not be cost effective to find the set of k vertices which minimize the cost of the new partition, the "mob" to be transfered is chosen to the k vertices which would give the k best partitions if transfered individually.

Because of the group transfer, this algorithm may be more efficient when done in parallel; however, because of the group transfer, one step may not produce as good a partition as one step of the Kernigan-Lin algorithm.

2.1.2 Spectral algorithms

The problem of finding a minimum cost partition can be expressed as an integer quadratic program. Let us assume for simplicity that the number of vertices of G is even and all node and edge weights are 1; thus, we need V_1 and V_2 to have the same size.

For each vertex i, let x_i be a ± 1 variable which is 1 if $i \in V_1$ and -1 if $i \in V_2$. Then the condition that there be an equal number of vertices in each of the two parts of the partition is expressed by the linear equation $\sum_i x_i = 0$.

We write $i \sim j$ to mean vertices i and j are adjacent; then the number of edges crossing the partition is given by $\frac{1}{8} \sum_{i \sim j} (x_i - x_j)^2$, since if i and j are in the same part, $x_i = x_j$, and so $(x_i - x_j)^2 = 0$, while if i and j are in different parts, $x_i \neq x_j$, so $x_i - x_j = \pm 2$ and $(x_i - x_j)^2 = 4$. Since each edge is counted twice (once each way), this accounts for the 1/8. Then we have

$$\sum_{i \sim j} (x_i - x_j)^2 = \sum_{i \sim j} (x_i + x_j)^2 - 2 \sum_{i \sim j} x_i x_j = \mathbf{x}^{\mathrm{T}} L \mathbf{x}$$

where L is the *Laplacian Matrix* of the graph, defined by $L = [l_{ij} \mid i, j \in V]$, where $l_{ii} =$ the degree of vertex i and for $i \neq j$, $l_{ij} = 0$ if no edge connects i and j, and $l_{ij} = -1$ if an edge connects i and j.[3]

Thus, we have the integer quadratic program

$$\text{Minimize } \mathbf{x}^{\mathrm{T}} L \mathbf{x}$$
$$\text{Subject to } \mathbf{1}^{\mathrm{T}} \mathbf{x} = 0 \tag{1}$$
$$x_i = \pm 1$$

where $\mathbf{1}$ is the vector of all 1s. However, integer programming is also NP-complete, so this does not help very much.

But, suppose we now replace the constraint that $x_i = \pm 1$ with the weaker linear constraint $\mathbf{x}^{\mathrm{T}} \mathbf{x} = n$. We are left with a quadratic program, which is easy to solve:

$$\text{Minimize } \mathbf{x}^{\mathrm{T}} L \mathbf{x}$$
$$\text{Subject to } \mathbf{1}^{\mathrm{T}} \mathbf{x} = 0 \tag{2}$$
$$\mathbf{x}^{\mathrm{T}} \mathbf{x} = n$$

This program can be solved particularly easily by finding the second smallest eigenvalue of L, as follows: Let the eigenvalues of L be $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ (this is called the spectrum of L). It can be shown that $\lambda_1 = 0$,

---

[3] If A is the ordinary adjacency matrix of the graph, and D the diagonal matrix of degrees, then $L = D - A$.

and if the graph is connected, then $\lambda_1 < \lambda_2$. Furthermore, it is easy to see that if $\mathbf{x}$ is an eigenvector with eigenvalue λ_2 (weighted such that $\mathbf{x}^T\mathbf{x} = 1$), then it solves the quadratic program (2). (See, for instance, [13] for details of all of this.)

The spectral method finds this second largest eigenvalue and a corresponding eigenvector, and then uses this solution to (2) to find an approximate solution to the integer program (1). For instance, it might take the vertices corresponding to the $n/2$ smallest components of the eigenvector and put them in V_2, while putting the remaining (largest) $n/2$ vertices in V_1.

For history of the spectral method, see [12]. It can be extended to partition in four or eight parts [13, 24] by using additional eigenvalues, but not beyond that; however, the spectral algorithm can be repeated on the resulting parts of the graph.

2.1.3 Multilevel Algorithms

The multilevel partitioning algorithm works by first "coarsening" the graph by merging some sets of nodes into new "supernodes", typically by shrinking edges. The new graph is then partitioned; this partition induces a partition of the original graph, which can be improved by the Kernigan-Lin algorithm. Good results are achieved by doing several rounds of coarsening until the coarsened graph is relatively small, then partitioning by some other method (for example, spectral). Several rounds of uncoarsening and Kernigan-Linning follow [14,16,17].

Currently, multilevel methods seem to be the best algorithms for partitioning graphs well in a short amount of time. (Simulated Annealing seems to partition graphs slightly better but in much more time.)

2.2 Work on Graph Embedding

There has been a certain amount of work on "embedding" a source graph into a target graph. In general, though, there are several differences between this work and our thinking. In an embedding, the two graphs are (roughly) of equal size, and the work has concentrated on finding maps between the vertex sets of the two graphs which have small "dilation" (the distance between the images of two adjacent vertices of the source graph in the target graph; we require dilation 1), "congestion" (the maximum number of edges of the source graph mapped to a single edge of the target graph), and similar parameters.

In addition, this work tends to focus on finding algorithms to map one class of graphs (*e.g.*, grids) onto another class of graphs (*e.g*, hypercubes) so as to always use the smallest possible member of the target class in such a way as to obtain the provably best embedding in terms of one or more of the parameters (*e.g.*, dilation no more than 2 [4]). See also [25].

This work does not appear to apply directly to our effort. First, the problems are very hard, so results are limited to special cases. Second, we do not wish to allow any dilation in our maps, but dilation of more than one is often required in an embedding. Finally, in general our guest graph is larger than the host graph, rather than the same size used in the embedding problem (though this problem could be eliminated by using a multilevel type algorithm, where the guest graph is coarsened to a graph with the same number of vertices

as the host graph). The work in [10], however, extends graph embedding techniques in the direction of our problem.

It should also be noted that the term "graph embedding" is more often used to refer to drawing a graph (in some surface) so that no edges cross [9]. This work appears to have less relevance to our problem, though see section 2.3.

2.3 Work on Our Problem

There has not been a lot of work on our more general problem. For embedding into hypercubes, recursive bisection often works well because of the structure of the graph. [13]'s quadrisections and octasections also can be interpreted as embedding into a small cube.

A simple technique to attack the problem is for each processor in the host graph to compute its load, globally average these, and then have each processor (somehow) adjust its load to the average. One method for adjusting the load is the "flow algorithm", discussed in my report from last summer [21].

In summary, the flow algorithm finds the amount of load that each processor must gain or lose to be at the global average, as above. These numbers are used as the input to a network "max-flow" algorithm: that is, finding local movements of the work load along edges of the host graph which will result in each processor's change in work load being satisfied. There are many algorithms for this classical problem.

The downside of such an algorithm is the serial dependency introduced by computing the global average (and running the flow algorithm). Serial dependencies can hamper parallel efficiency, in that most processors will be idle during the time the serial computation is running. Also, as the host graph gets larger, the serial time required here would increase.

A straight-forward way to get around this problem is for each processor to average its load with its neighbors; call this *local averaging*. This, however, has some problems of its own. Either one must accept less uniformity of the load or repeat the algorithm multiple times; in the latter case, it might be faster to do the global computation.

Heirich [11] points out another problem. Consider the case where the host graph is 1-dimensional grid, so processor p_i is connected to p_{i-1} and p_{i+1} (except at the ends). Then (again, except at the ends) the previous algorithm will arrive at a steady state where the load u_i on processor p_i satisfies $u_i = \frac{1}{2}(u_{i-1} + u_{i+1})$. It is shown that this means u_i is a second-order solution to the Laplace partial differential equation $\frac{\partial^2 u_i}{\partial i^2} = 0$. The solutions to this PDE are $u_i = mi + b$, but the solution we are after, load balance, requires also that $m = 0$. (It should be noted that in real architectures, the presence of boundaries or wrap arounds will eventually eliminate this problem.)

The previous analysis suggested [11]'s algorithm, "Diffusion". The basic idea is replace the Laplace equation with an equation for heat flow, $\frac{\partial \mathbf{u}}{\partial t} - \nabla^2 \mathbf{u} = 0$. An iterative solution to this equation is used as the load balancing algorithm. There are two drawbacks: the algorithm is quite complicated compared with local averaging, and it appears that we are restricted to grid-like graphs as host graphs (since otherwise it is hard to make sense of ∇). This last is probably not a big restriction for the case of PIC plasma simulations, as

these graphs are indeed grid-like. (And it might be possible to embed (in the sense of [9]) the host graph into a surface and interpret the differentials there.)

3   The Test Bed

The major effort of the time I spent at MHPCC was writing the foundation of a test bed program. The test bed consists of C++ code which implements a distributed graph class on any parallel processor which implements PVM. The code is written in a literate programming style using a new literate programming tool called Pig. It currently runs on a couple of PCs runing Linux as well as on the SP2 at MHPCC.

3.1   Hardware Systems Used

Our code was written on a pair of 80486 PC computers running the Linux [27] operating system at Philips Labs. It was later ported to the SP-2 multiprocessor at the Maui High Performance Computing Center. Most recently, the code moved to a PC, also running Linux, at Texas A&M University-Kingsville.

3.2   Software Tools Used

C++ [6] is an object oriented extension of the Cprogramming language. C++ was chosen as an implementation language because, thanks to the GNU C++ compiler, it is relatively portable; it is at least somewhat object oriented; and can easily call C libraries such as PVM. In addition, I already know C++, and am reasonably productive writing code in it. Objective C might have been a better choice, given some of the problems encountered, but it is much less available that C++; in particular, the GNU Objective C compiler is much less mature than the C++ compiler.

PVM is one of many libraries of C functions which provide message passing facilities, both on MPPs and on networks of workstations. PVM was chosen because it is relatively portable, and I already knew it. PVM is nicer in some respects than the proposed Message Passing Protocol (MPP), but on the SP2 at MHPCC, using MPP will eventually be faster. Since the PVM code is isolated in a few classes, it will be (hopefully) not too hard replace PVM by MPP or another library if this is needed for performance.

Pig is an experimental literate programming system written in Perl [26] by myself. In retrospect, it would have been better to use one of the publicly available systems such as noweb [22] or nuweb [3]; eventually I expect to convert over to some such system.

The main "innovation" in Pig is that the source code and the "woven" output (to be sent to LaTeX) are one and the same; other systems use a separate "weave" program to prepare the woven output. We eliminate the need for a separate program by making the "code chunks" into LaTeX environments. The gain, however, seems small compared with the problems of using the non-standard, immature Pig program; in addition, the weave programs often proved additional features (such as indexing and easy formatting of code chunks embedded in paragraphs) which are difficult to obtain in LaTeX macros.

3.3 Structure of the Test Bed Classes

Rather than just call PVM routines, we hide the calls in C++ classes. As noted above, this means that it will be relatively easy to replace PVM by another library later.

The current structure of the class hierarchy is described below; in the long term, I hope to extend this collection of classes to a general purpose message passing system for C++.

At the lowest level, classes represent *tasks* (or, rather, task handles), *network buffers* used to store messages before they are sent, and *messages*. In the later case, we really describe a class interface (by use of a virtual base class [5]) which all classes which will be sent as messages inherit from. Messages can be put in network buffers, which in turn can be sent to other tasks, or messages can be sent directly. At this level, the receiving task must explicitly receive the message(s) which are sent to it.

This level is built on standard C++ facilities and the PVM library; it also provides a data type (called a "Qid", short for uniQue IDentifier) which allows for the unique identification of objects over all tasks.

Built on the previous level and a little PVM is a class representing *active messages*, that is, messages which will execute some code on arrival. Because of the structure of C++ and PVM, the receiving program must call a special receive routine periodically to allow the active message to be received.

The distributed graph class represents a graph the vertices of which are distributed among several tasks. Each task has information about the vertices it "owns" along with all neighboring vertices. This information is stored in a LEDA graph on each vertex with some additional information. For instance, each vertex is marked as either local or remote, and the remote vertices keep track of which task owns the vertex in question.

LEDA [19] is a library of efficient data types and algorithms for C++ from Germany; it includes a graph class. I chose to build the distributed graph class on LEDA because it is freely available for research, generally seems to work okay, and saved me an enormous effort writing my own graph class.

The distributed graph class implements adding and deleting vertices and edges and moving edges between tasks. We expect to extend many of LEDA's graph algorithms to the distributed graphs as needed.

Currently, we have only one algorithm written for distributed graphs: a Kernigan-Lin algorithm between two tasks.

4 Future Plans

I intend to use the test bed, both this fall on an AFOSR followup grant, and in the future, to test the algorithms mentioned above (include my own flow algorithm) on a variety of graphs; I hope this will include some graphs derived from PIC plasma simulations. In addition to simulated load variations, eventually we hope to add a simple PIC simulation to give more realistic load variations. I am particularly interested in whether global algorithms like the flow algorithm actually out-perform local algorithms like local averaging and diffusion; if so, then there are tradeoffs to consider.

In the theoretical direction, recently there has been some work done on showing when problems like ours have "good" approximate algorithms. This work is related to the theory of NP-completeness. I would like to investigate whether these methods could be used to either find a (theoretically) good approximation algorithm

for the remapping problem, or alternatively to show that there are no such algorithms. Just as showing the remapping problem was NP-complete was helpful (in that it suggested that a polynomial-time exact algorithm was impossible), this might help in the search for good algorithms for remapping.

5 Bibliography

[1] S. T. Barnard and H. D. Simon, *Fast Multispectral Implementation of Recursive Spectral Bisection for Partitioning Unstructured Problems*, Concurrency 6 (1994), 101–117.

[2] C. K. Birdsall and A. B. Langdon, *Plasma Physics via Computer Simulation*, Adam Hilger, New York, 1991.

[3] P. Briggs, *Nuweb Version 0.87b: A Simple Literate Programming Tool*, distributed with nuweb, ND.

[4] M. Y. Chan and F. Y. L. Chin, *On Embedding Rectangular Grids in Hypercubes*, IEEE Transacations on Computers 37 (1988), 1285–1288.

[5] J. O. Coplien, *Advanced C++ Programming Styles and Idioms*, Addision-Wesley, Reading, MA, 1992.

[6] M. A. Ellis and B. Stroustrup, *The Annotated C++ Reference Manual*, Addison-Wesley, Reading, MA, 1990.

[7] C. M. Fiduccia and R. M. Mattheyses, *A Linear-Time Heuristic for Improving Network Partitions*, in 19th IEEE Design Automation Conference, 1982, 175–181.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability*, Freeman, New York, 1979.

[9] J. L. Gross and T. W. Tucker, *Topological Graph Theory*, Wiley Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, New York, 1987.

[10] A. K. Gupta and S. E. Hambrusch, *Load Balanced Tree Embeddings*, Parallel Computing 18 (1992), 595–614.

[11] A. Heirich, *Scalable Load Balancing by Diffusion*, Caltech, preprint, 1994.

[12] B. Hendrickson and R. Leland, *An Improved Spectral Graph Partitioning Algorithm for Mapping Parallel Computations*, Sandia National Laboratories, SAND92-1460, Albuquerque, NM, 1992.

[13] _____, *Multidimensional Spectral Load Balancing*, Sandia National Laboratories, SAND93-0074, Albuquerque, NM, 1993.

[14] _____, *A Multilevel Algorithm for Partitioning Graphs*, Sandia National Laboratories, SAND93-1301, Albuquerque, NM, 1993.

[15] R. M. Karp, *Reducibility Amoung Combinatorial Problems*, in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, 85–103.

[16] G. Karypis and V. Kumar, *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs*, University of Minnesota, Department of Computer Science, 95-035, Minneapolis, 1995.

[17] _____, *Parallel Multilevel Graph Partitioning*, University of Minnesota, Department of Computer Science, 95-036, Minneapolis, 1995.

[18] B. W. Kernigan and S. Lin, *An Efficient Heuristic Procedure for Partitioning Graphs*, Bell Systems Technical Journal 49 (1970), 291–307.

[19] S. Näher, *LEDA User Manual Version 3.0*.

[20] A. Pothen, H. D. Simon and K-P. Liou, *Partioning Sparse Matrices with Eigenvectors of Graphs*, SIAM Journal of Matrix Analysis and Applications 11 (1990), 430–452.

[21] M. Purtill, *Static and Dynamic Graph Embedding for Parallel Programming*, Phillips Laboratory, Final Report for AFOSR Summer Faculty Research Program, 1994.

[22] N. Ramsey, *Literate-Programming Can Be Simple and Extensible*, Department of Computer Science, Princeton University, distributed with noweb, 1993.

[23] J. E. Savage and M. G. Wloka, *Parallelism in Graph-Partitioning*, Journal of Parallel and Distributed Computing 13 (1991), 257–272.

[24] P. R. Suaris and G. Kedem, *An Algorithm for Quadrisection and Its Application to Standard Cell Placement*, IEEE Transactions on Circuits and Systems 35 (1988), 294–303.

[25] A. S. Wagner, *Embedding All Binary Trees in the Hypercube*, Journal of Parallel and Distributed Computing 18 (1993), 33–43.

[26] L. Wall and R. L. Schwartz, *Programming Perl*, O'Reilly & Associates, Sebastopol, CA, 1991.

[27] M. Welsh and L. Kaufman, *Running Linux*, O'Reilly & Associates, Sevastopol, CA, 1995.

USING LASER DYNAMICS TO PROBE INTRACAVITY COLLISIONAL RELAXATIONS

Jeff Nicholson

Graduate Student

Wolfgang Rudolph

Associate Professor

Department of Physics and Astronomy

The University of New Mexico

Albuquerque, NM 87131

Report for:

Graduate Student Summer Research Program

Phillips Laboratory

September 1995

USING LASER DYNAMICS TO PROBE INTRACAVITY COLLISIONAL RELAXATIONS

Jeff Nicholson

Graduate Student

Wolfgang Rudolph

Associate Professor

Department of Physics and Astronomy

The University of New Mexico

Albuquerque, NM 87131

Abstract

The pulse parameters of a photolytic, gain-switched iodine laser are measured for different pressures of active gas and for various buffer gases. A rate equation model of the laser dynamics is fit to experimental data, to obtain information about collision-controlled relaxation processes of atomic iodine.

USING LASER DYNAMICS TO PROBE INTRACAVITY COLLISIONAL RELAXATIONS

Jeff Nicholson and Wolfgang Rudolph

1 Introduction

Photolysis of a molecule by a short laser pulse results in a nonthermal velocity distribution. Due to collisions changing the direction and the magnitude of the velocity vector the initial distribution relaxes to an isotropic Maxwellian distribution. It was found that the velocity relaxation comprises two components — a relaxation of the speed to a Maxwellian distribution at a certain equilibrium temperature and a relaxation of the direction of the velocity vector to an isotropic distribution. Being controlled by collisions, the relaxation rates of these two processes can be different and depend on the collision partners (photolytic products and buffer gas). Heavy collision partners tend to produce a directional relaxation which can be several times faster than the corresponding time constant for the speed [1]. While this effect follows from hard-sphere collision theory, the effect of the structure (e.g., mono-atomic noble gas versus polar multi-atomic molecule) is much more complex and not very well understood. The latter is due partly to the lack of suitable experimental data.

Information about the velocity can be gained by measuring the spectral absorption profile of one of the products of the photolysis as a function of time. A corresponding gain measurement in such a pump-probe experiment can be performed if one of the reaction products emerges in an excited state. A number of such experiments have been executed in the past [2], [3] and yielded collision rates as a function of collision partner (buffer gas) and the temporal behavior of the velocity profile as it developed from a nonequilibrium to an equilibrium distribution. A typical measurement required a narrow bandwidth probe laser and a time resolving data acquisition system such as a boxcar integrator or a digital storage oscilloscope.

Recently the temporal shape of a gain switched laser pulse was used to derive relaxation parameters of the active gas after short pulse photolysis [4]. The pump pulse, being shorter than all relaxation processes of interest here, prepares the sample in a certain state. The time evolution of the gain coefficient can be probed by the evolving laser pulse itself provided the cavity lifetime is sufficiently short. This technique with its high signal to noise ratio, in principle, allows determination of the velocity relaxation parameters after a single photolysis event.

Immediately after photolysis off CF_3I with a short (20 ns) pump pulse at 248 nm propagating along the laser axis the gain profile differs considerably from a Doppler profile at room temperature. Since the pump photon energy is much greater than the energy needed for dissociation the mean energy of the photofragments corresponds to a higher temperature and larger mean velocity then when at room temperature. Since the pump laser is polarized perpendicularly to the resonator axis the nascent velocity distribution is peaked in

$F'=3$ +0.265 cm^{-1}

$^2P_{1/2}$

7603.15 cm^{-1}

2   -0.371 cm^{-1}

F=4   +0.113 cm^{-1}

$^2P_{3/2}$ 3   -0.028 cm^{-1}
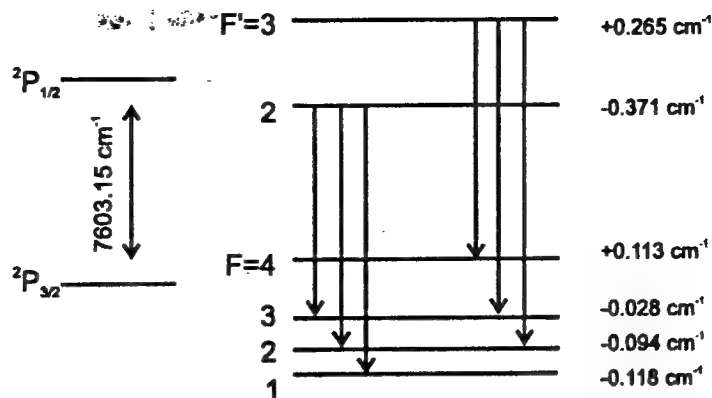
2   -0.094 cm^{-1}

1   -0.118 cm^{-1}

Figure 1: Energy levels and allowed transitions of atomic iodine

this direction making the gain at resonance greater as compared to an isotropic distribution. At the same time, however, the speed distribution is broader than at thermal equilibrium (room temperature). Therefore, the net gain at resonance is actually smaller as compared with the equilibrium situation.

While the laser radiation develops from spontaneous emission, the gain media cools thus narrowing the Doppler profile and increasing the gain available to the laser. At a certain intensity of the laser the gain at resonance is depleted, burning a hole in the velocity distribution, and leading to a decrease in laser power. Because of cross-relaxation processes filling the hole, the gain and thus the laser output can increase again which manifests itself as a distinct shoulder or a second pulse.

We will show in detail how the various relaxation times can be obtained for different buffer gases and pressures from fitting the experimental laser pulse profiles with a comprehensive rate equation model of the laser.

2 Experimental

When pumped at 248 nm, CF_3I dissociates with approximately 92% efficiency into excited state iodine (I*) and CF_3. The other 8% of the iodine goes into the ground state [5]. The energy levels of atomic iodine are shown in Fig. 1. The strongest transition and the one that lases under typical experimental conditions is the F=3 to F=4 transition corresponding to a wavelength of 1.315 μm.

The experiment consisted of a photolytic iodine laser pumped by a KrF laser (see Fig 2). The gain media used was 0.8 torr CF_3I in an 85 cm cell with fused silica windows at Brewster's angle for 1.3 μm. The cell was pumped longitudinally by a 248 nm pulse with a width of approximately 20 ns and an energy of 200 mJ. The pump beam was focused to a 10mm by 4mm spot in the cell. At 0.8 torr of CF_3I with these pumping
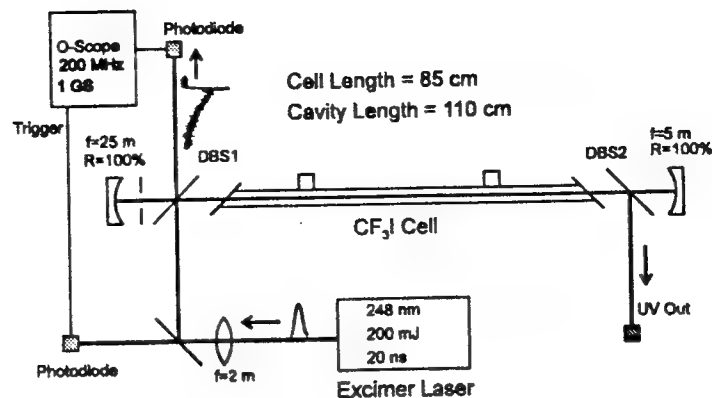
Figure 2: Schematic diagram of the photolytic, gain-switched iodine laser used to study collisional relaxation processes.

conditions, 15% of the CF_3I was dissociated into CF_3 plus I^*.

In addition to the CF_3I in the cell, buffer gases were introduced into the gain media to change collision times without changing the number of active molecules. Buffer gases used were He, Ne, Kr, and Xe. CO_2 and N_2 were also used in order to investigate the difference between the hard sphere collisions of noble gases and inelastic collisions of polyatomic molecules.

The cavity consisted of two 100% reflecting mirrors separated by 1.1m, with radius of 5 m and 25 m. The pump beam was coupled into the cavity with a dichroic beam splitter and coupled out with a second dichroic beam splitter. The output of the iodine laser was taken from a small reflection off the first dichroic beam splitter, while the second beam splitter was set at Brewster's angle for 1.3 μm. Iodine pulse energies were on the order of several μJ.

After laser action, only about 90% of the molecular fragments recombine to form CF_3I, the rest of the fragments going into processes such as dimerization [5]. I_2 is a strong quencher of the $^2P_{1/2}$ state, leading to a degradation of the iodine laser output after multiple shots of the excimer laser. Therefore, the iodine cell was refilled after every fourth or fifth excimer pulse. Note, that while these processes proved to be important after multiple shots, they do not play a significant role in single shot pulse dynamics on a time scale < 10 μs because of their slow (e. g. $\sim 4 * 10^{-31} cm^6 mol^{-2} s^{-1}$ for dimerization) rate.

The iodine pulse was detected with a fast (< 500 ps risetime) InGaAs photodiode and viewed on a digital oscilloscope (Textronix TDS 350). A part of the excimer pulse was also incident on a photodiode. This signal was used to trigger the oscilloscope and give an absolute measure of the delay between the arrival of the excimer pulse and onset of the iodine pulse.

A typical iodine pulse with 0.8 torr CF_3I in the gain cell is shown in Fig. 3. The excimer pulse occurs

35-5
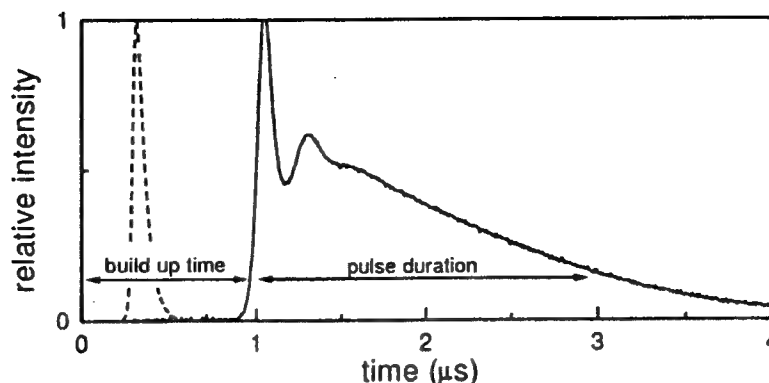
Figure 3: A typical pulse from a low pressure photolytic iodine laser. CF_3I pressure was 0.8 torr. The dotted line is the result from a simple rate equation model without velocity relaxation mechanisms. The pump pulse occurs at t=0.

at $t = 0$, and is essentially a delta function on this time scale. The dotted pulse is the result obtained from a simple three level rate equation model that does not include any velocity relaxations. The experimental pulse shape has a much longer delay between pump pulse and onset of lasing, a much slower decay time, and a distinct second peak. All of these features are the result of the finite relaxation times for the velocity distribution. Fig. 3 also shows how the build up time and pulse duration were defined. These quantities were measured at 10% of the pulse peak intensity.

Simultaneous lasing of multiple longitudinal modes is frequently observed in gain switched iodine lasers [6],[7]. Fig. 4 shows two iodine pulses taken under identical experimental conditions that show clear longitudinal mode beating. However, the relative strengths and build up times are different for the two modes for the two pulses, obscuring or enhancing the pulse features shown Fig 3. Therefore, an intracavity aperture was used to force the laser to oscillate both with a single longitudinal mode and in TEM_{00} mode. An aperture diameter of less than 2 mm was used as compared to a TEM_{00} waist of 2.2 mm. This eliminated mode beating in approximately 50% of the iodine pulses. The aperture also forced the laser frequency to be much less than half the free spectral range of the cavity (70 MHz) away from line center.

Although the 3-4 transition is the strongest transition in atomic iodine, under certain experimental conditions both the 3-4 and 2-2 transition have lased [8],[9]. In order to be sure only the 3-4 transition was lasing, a streak camera (resolution < 5 ps) was used to look for the fast modulation that would result from simultaneous lasing and beating of two hyperfine modes. This mode beating would have a frequency of 13 GHz, the difference frequency between the 2-2 and 3-4 transitions. However, there was no evidence of a 13 GHz signal, in any of the measurements made with the streak camera.

In addition, the rate equation model, described in Section 3.2 was used to investigate the possibility of
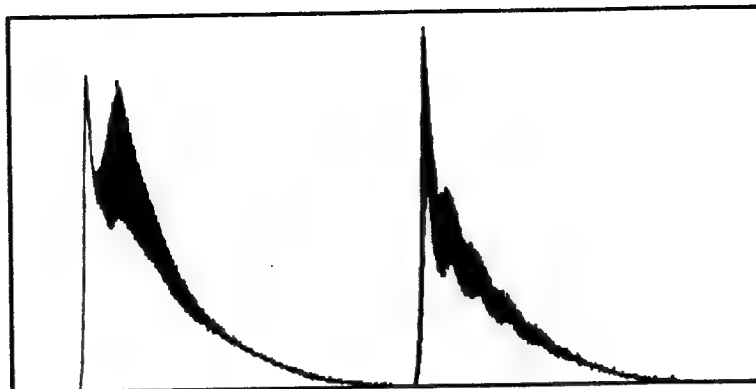
Figure 4: Two examples of iodine pulses with clear longitudinal mode beating
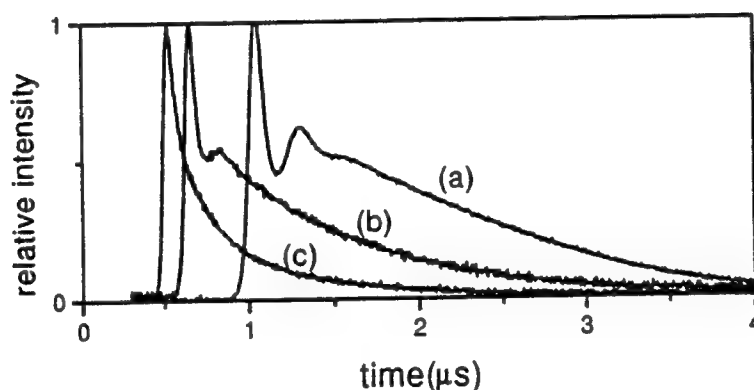
Figure 5: The change in pulse shape as Xe is added to the gain cell. 0.8 torr CF_3I + (a) 0 torr Xe, (b) 0.8 torr Xe, (c) 2.4 torr Xe. The peak intensities have been normalized to one for comparison purposes

the 2-2 line lasing. The cross section of the 3-4 transition is twice as large as the 2-2 transition. When only the 2-2 transition is included in the model, with the 3-4 transition artificially suppressed, the transition is able to lase, but the build up time is 2 μs greater and the peak intensity is one fourth of the values obtained when only the 3-4 line is included in the model. However, when both (single longitudinal mode) hyperfine transitions are included in the model, only the 3-4 line is able to lase. The 3-4 line builds up intensity quickly compared to the 2-2 line, and depletes the F=3 level. Then a fast population transfer due to collisions with ground state iodine [10] equalizes the upper hyperfine levels. This depletes the population in the upper level of the 2-2 line, preventing it from lasing. From this, we concluded that the experimental iodine pulses contained the 3-4 line only.

Fig. 5 shows the change in pulse shape as a buffer gas (in this case, Xe) is introduced into the gain cell.

As the buffer gas' partial pressure is increased, the build up time of the pulse decreases, the second peak disappears, and the pulse width decreases. The pulse energy remains approximately constant with increasing buffer gas pressure.

A summary of the effects of buffer gas on pulse parameters is shown in Fig. 6. Because the number of active molecules remains constant, the change in pulse shapes can be attributed to changing the collisional dynamics caused by adding buffer gas. Fig. 6(a) shows Noble gases; Fig. 6(b) shows polyatomic molecules. Xe is included in (b) for comparison purposes.

As the partial pressure of buffer gas in the gain cell increases, collision rates increase, and build up time and pulse duration decrease, and the second peak becomes smaller and disappears. However, more massive buffer gases are more effective than lighter buffer gases in thermalizing the velocity distribution. In fact, the changes in the velocity relaxation times caused by helium are too small to account for the observed changes in the pulse shape, but can be explained by the small increase in the homogeneous linewidth caused by pressure broadening (see Section 4). The increase in homogeneous linewidth is insufficient to explain the changes in pulse shape caused by the more massive buffer gases, however.

3 Theoretical Model

3.1 Collision Dynamics

Immediately after photodissociation, the velocity distribution of the photofragments is separable [11]; it can be written as a product of a function of speed, f and an angular function, I,

$$F(\vec{v}, t) = f(v, t) I(\theta, \phi, t) \tag{1}$$

where the angular distribution function $I(\theta, \phi, t)$ has the form

$$I(\theta, \phi, t) = \frac{1}{4\pi} \left[1 + \beta(t) \ P_2(\cos \theta) \right]. \tag{2}$$

β is the anisotropy parameter, and θ and ϕ are the angles between the fragment's velocity vector and the electric field vector of the dissociating beam. Depending on whether the molecule undergoes a parallel or perpendicular transition during dissociation, β can take the value 2 or -1 respectively. CF_3I dissociating into CF_3 and I* is a parallel transition with $\beta = 1.4$ [12].

In previous Doppler spectroscopy experiments, a probe beam was used to measure the Doppler profile. In general the propagation direction of the probe beam is different from that of the dissociating beam. However, in this experiment, the probe beam is the laser field, and because the cell is longitudinally pumped, the pump and probe beams are parallel. The relevant coordinate axis in relation to the resonator axis is shown in Fig. 7. Immediately after photodissociation, the I* atoms tend to move perpendicular to the laser axis.
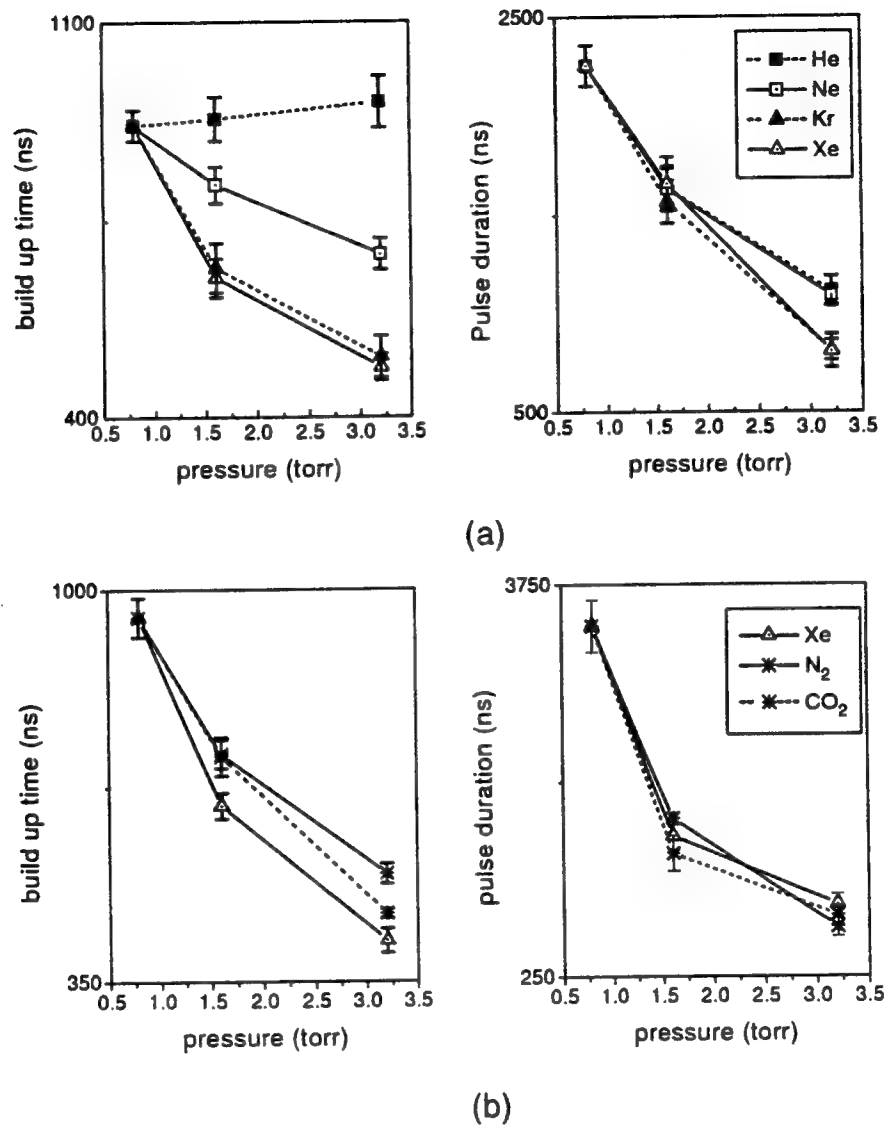
Figure 6: Summary of the effects of the addition of buffer gas on the build up time and pulse duration. (a) Noble gases. (b) Polyatomic molecules.
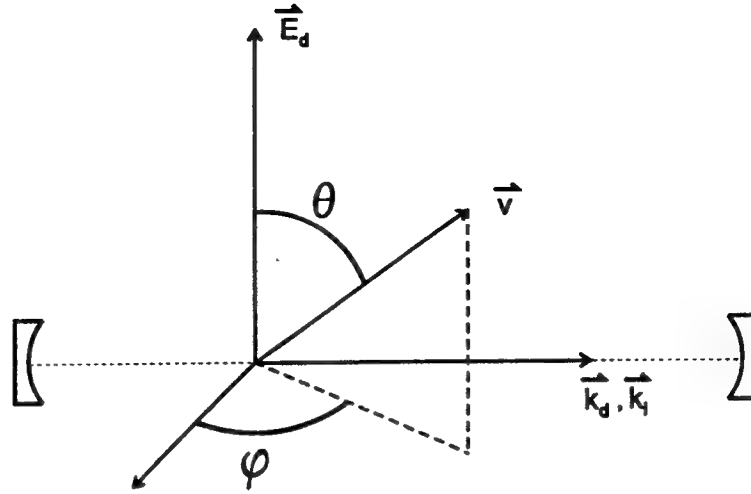
Figure 7: Coordinate axis in relation to the iodine resonator axis. The subscript d refers to the dissociating beam, and l refers to the iodine laser beam.

It was recently shown that as $F(\vec{v}, t)$ relaxes to thermal equilibrium, it remains approximately separable [13]. The mean velocity relaxes exponentially to its equilibrium value with a time constant τ_v

$$\bar{v}(t) = \bar{v}_{\text{fin}} + (\bar{v}_{\text{in}} - \bar{v}_{\text{fin}})e^{-t/\tau_v} \tag{3}$$

where $\bar{v}_{\text{in}}$ and $\bar{v}_{\text{fin}}$ are the initial and final mean velocities respectively. β relaxes to zero (isotropy) with a time constant τ_b

$$\beta(t) = \beta_{\text{in}} e^{-t/\tau_b} \tag{4}$$

A full analysis using the Boltzmann collision operator shows that the ratio τ_v/τ_b becomes larger for more massive collision partners [1]. The approximate separability and the relative ratios of the time constants has been experimentally verified for a few examples using dissociation of C_3F_7I at 266 nm [2].

Stimulated emission has a much smaller effect then collisions on the separability of the velocity distribution. If we assume, for the moment, that the laser field is single frequency, and there is no homogeneous broadening of the laser transition, then only atoms moving perpendicular to the laser axis are removed from the velocity distribution of the excited state atoms. If we begin with a separable velocity distribution that has the form of Eq. (1) then change in the velocity distribution due to stimulated emission can be written as

$$f(v,t)I(\theta,\phi,t) \rightarrow f(v,t)(1 - a\delta(\theta))I(\theta,\phi,t) = f(v,t)I'(\theta,\phi,t) \tag{5}$$

where a is the fraction of total atoms removed from the excited state due to stimulated emission. $F(\vec{v}, t)$ is still separable, although I is no longer of the form of Eq. (2). Although this is strictly true for a monochromatic
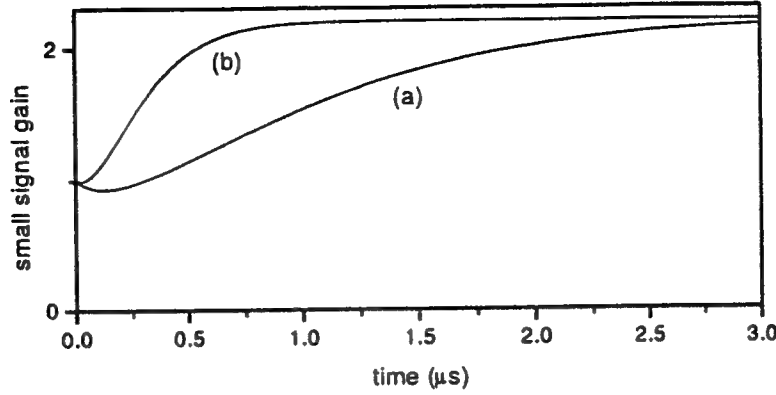
Figure 8: Small signal gain at line center. (a) $\tau_v = 650$ns, $\tau_b = 120$ns (b) $\tau_v = 180$ns, $\tau_b = 60$ns

laser field and no homogeneous broadening, the linewidth of the laser and the homogeneous linewidth are so much smaller than the inhomogeneous linewidth ($\approx$ 10 MHz vs 1 GHz) that it remains a good assumption. Thus, the dominant reasons for the loss of separability remain those discussed in Ref. [2], such as collisions between I^* and buffer gas.

The Doppler profile is obtained from the velocity distribution by summing all the fragments that have the same velocity component in the direction of the probe beam

$$\Gamma_D(w, \theta', \phi', t) = \int d^3 v \ F(\vec{v}, t) \ \delta(w - \vec{v} \cdot \vec{k}_p) \tag{6}$$

where w is the velocity component in the direction of the probe beam. It is related to the Doppler frequency shift by $\nu - \nu_0 = \nu_0 \frac{w}{c}$ where ν_0 is the center frequency of the transition and c is the speed of light. The primed angles refer to the probe direction. For the particular case where the pump and probe are parallel, using Eqs (1) and (2), Eq. (6) reduces to [3]

$$\Gamma_D(\nu - \nu_0, t) = \frac{c}{\nu_0} \int_{|\frac{c}{\nu_0}(\nu - \nu_0)|}^{\infty} \frac{v}{2} \left[1 - \frac{\beta(t)}{4} \left(3 \left(\frac{c(\nu - \nu_0)}{\nu_0 v} \right)^2 - 1 \right) \right] f(v, t) \ dv. \tag{7}$$

From here on, we write Γ_D as a function of ν rather than w. Eq. (7) can be solved analytically for line center ($\nu = \nu_0$), assuming a Maxwellian velocity distribution for $f(v, t)$:

$$\Gamma_D(0, t) = \frac{2c}{\nu_0 \pi} \left(1 + \frac{\beta(t)}{4} \right) \frac{1}{\bar{v}(t)}. \tag{8}$$

A plot of Eq. (8) is shown in Fig. 8 for two values of τ_v and τ_b. If we neglect for the moment the small amount of ground state iodine produced after dissociation, $\Gamma_D(0, t)$ is proportional to the small signal gain at line center. This clearly shows that as the relaxation times decrease, the gain available to the iodine laser reaches equilibrium faster, shortening the build up time of the iodine pulse. In addition, the build-up time of the laser pulse is due primarily to τ_v and only to a lesser extent τ_b.

35-11

In order to connect the relaxation of the Doppler profile $\Gamma_D(\nu - \nu_0, t)$ to the relaxation of the velocity distribution $F(\vec{v}, t)$, take the time derivative of Eq. (6)

$$\frac{d\Gamma_D(\nu - \nu_0, t)}{dt} = \int d^3v \left[\frac{dI(\theta, \phi, t)}{dt} f(v, t) + \frac{df(v, t)}{dt} I(\theta, \phi, t) \right] \delta\left((\nu - \nu_0) - \frac{\nu_0}{c} \vec{v} \cdot \vec{k}_p \right). \tag{9}$$

Therefore, for small enough time steps, the contribution to the time rate of change of $\Gamma_D(\nu - \nu_0)$ can be separated into two terms. One term controls how $\Gamma_D(\nu - \nu_0, t)$ changes due to a relaxation of the angular anisotropy function $I(\theta, \phi, t)$, and the other term controls how $\Gamma_D(\nu - \nu_0, t)$ changes due to a relaxation of the speed distribution $f(v, t)$.

The integral that arises from the first term in the brackets in Eq. (9) describes how the Doppler profile returns to equilibrium due to a relaxation of the anisotropy only. If $I(\theta, \phi, t)$ relaxes to equilibrium exponentially, (as is the case for the anisotropy function in Eq. (2)), then the relaxation of $\Gamma_D(\nu - \nu_0, t)$ to equilibrium due to the anisotropy relaxation in $F(\vec{v}, t)$ can be written

$$\frac{d\Gamma_D(\nu - \nu_0, t)}{dt} = \frac{\Gamma_D^{eq}(\nu - \nu_0, t) - \Gamma_D(\nu - \nu_0, t)}{\tau_b}. \tag{10}$$

$\Gamma_D^{eq}(\nu - \nu_0, t)$ is the Doppler profile whose velocity distribution has the same mean velocity as $\Gamma_D(\nu - \nu_0, t)$ but no anisotropy; this is just a Gaussian profile with a (time dependent) full width at half maximum $\Delta\nu_D(t)$. $\Delta\nu_D(t)$ is simply related to the mean velocity by $\Delta\nu_D(t) = \nu_0\sqrt{\pi \ln 2}/c\bar{v}(t)$. Because the Doppler width is proportional to the mean velocity, $\Delta\nu_D$ also relaxes exponentially with the time constant τ_v

$$\Delta\nu_D(t) = \Delta\nu_{\text{fin}} + (\Delta\nu_{\text{in}} - \Delta\nu_{\text{fin}})e^{-t/\tau_v} \tag{11}$$

where $\Delta\nu_{\text{in}}$ and $\Delta\nu_{\text{fin}}$ are the initial and final Doppler width respectively.

Eq. (10) is sufficiently general to describe any relaxation of an anisotropy in the velocity distribution. It will serve to relax the initial angular anisotropy caused by photodissociation of CF_3I. In addition, stimulated emission from radiation at the center of the inhomogeneously broadened transition will remove atoms moving perpendicular to the laser axis and burn a hole in the Doppler profile. Eq. (10) will also relax such an induced angular anisotropy caused by the build up of laser radiation.

The second term in Eq. (9) shows how the Doppler profile relaxes due to the cooling of the velocity distribution. If we assume that immediately after dissociation, $f(v, t)$ is a Maxwellian distribution with a high temperature, $f(v, t)$ remains a Maxwellian distribution as the gas cools.

$$f(v, t) \propto \exp\left(-\frac{4v^2}{\pi\bar{v}^2(t)} \right) \tag{12}$$

Then, at time $t + dt$, $f(v, t + dt)$ is

$$f(v, t + dt) \propto f\left(\frac{\bar{v}(t)}{\bar{v}(t + dt)} v, t \right) \tag{13}$$
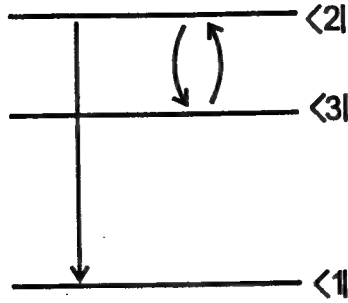
Figure 9: Diagram for levels and transitions used in the rate equation model. The laser transition is from $\langle 2|$ to $\langle 1|$. Population in $\langle 2|$ and $\langle 3|$ is equalized through collisions with population in $\langle 1|$.

Using Eq (6) and keeping $I(\theta, \phi, t)$ constant for small dt, we can relate the Doppler profile at time $t + dt$ to the Doppler profile at time t by

$$\Gamma_D(\nu - \nu_0, t + dt) = \alpha(t)\Gamma_D\left(\frac{\Delta\nu_D(t)}{\Delta\nu_D(t + dt)}(\nu - \nu_0), t\right) \tag{14}$$

where the proportionality constant $\alpha(t)$ is used to insure

$$\int \Gamma_D(\nu - \nu_0, t)d\nu = 1 \tag{15}$$

Therefore the functional form of $\Gamma_D(\nu - \nu_0, t)$ does not change its shape; only the width of the function changes. This gives a method for determining how $\Gamma_D(\nu - \nu_0, t)$ changes due to relaxation of $\Delta\nu_D(t)$.

3.2 Laser Rate Equations

The energy level diagram used for the laser model is shown in Fig. 9. Because of the fast relaxation processes between the ground state hyperfine levels (time constant < 1 ns [8]) we lump the lower levels shown in Fig. 1 into an effective ground state $\langle 1|$. Lasing can then occur between $\langle 2|$ and $\langle 3|$ and this ground state. The result of the photolysis is a population in the two excited states of iodine distributed according to the degeneracies of the states [10]. There is a relaxation process involving collisions with unexcited atomic iodine between levels $\langle 2|$ and $\langle 3|$ with a rate constant $k_{2,3} \approx 1.5 * 10^{-9}\text{cm}^3\text{mol}^{-1}\text{s}^{-1}$ [10] which tends to equalize the population once an asymmetry has been created:

$$\frac{d}{dt}N_2 = k_{2,3}N_1\left(N_3 - \frac{g_3}{g_2}N_2\right) \tag{16}$$

where N_i and g_i ($i = 1, 2, 3$) is the occupation number and degeneracy, respectively of level i, and $N_2 + N_3$ is the total number density of excited iodine. Because of this cross relaxation and the two times larger interaction cross-section, under typical conditions used in our experiments, lasing occurs only between states $\langle 2|$ and $\langle 1|$.

To formulate the rate equations for the laser we assume that only one single longitudinal mode oscillates and is in resonance with the center of the inhomogeneously broadened gain transition $\langle 2| \rightarrow \langle 1|$ at a frequency $\nu = \nu_0$. The laser spectrum is supposed to be (much) smaller than the homogeneous line width $\Delta\nu_h$.

Let us define occupation numbers of a subensemble of particles with a resonance frequency ν per frequency interval $[\nu - \frac{d\nu}{2}, \nu + \frac{d\nu}{2}]$, $\bar{N}_i(\nu - \nu_0)$, such that

$$\bar{N}_i(\nu) = \Gamma_D(\nu - \nu_0)\, N_i \tag{17}$$

and

$$N_i = \int_0^\infty \bar{N}_i(\nu)\, d\nu \tag{18}$$

The rate equations for these occupation numbers, including relaxation between the upper hyperfine levels, and relaxation of the velocity distribution anisotropy as described in Eq. (10), can be written as:

$$\frac{d}{dt}\bar{N}_2 = -q_{21}\Gamma_h(\nu - \nu_0)F\left(\bar{N}_2 - \frac{g_2}{g_1}\bar{N}_1\right) + \frac{1}{\tau_b}\left[\bar{N}_2^{eq} - \bar{N}_2\right] \tag{19}$$

$$+\Gamma_D(\nu - \nu_0)k_{23}N_1\left(N_3 - \frac{g_3}{g_2}N_2\right)$$

$$\frac{d}{dt}\bar{N}_1 = q_{21}\Gamma_h(\nu - \nu_0)F\left(\bar{N}_2 - \frac{g_2}{g_1}\bar{N}_1\right) + \frac{1}{\tau_b}\left[\bar{N}_1^{eq} - \bar{N}_1\right] \tag{20}$$

$$\frac{d}{dt}N_3 = -\Gamma_D(\nu - \nu_0)k_{23}N_1\left(N_3 - \frac{g_3}{g_2}N_2\right) \tag{21}$$

$$\tag{22}$$

where

$$\Gamma_h(\nu - \nu_0) = \frac{\Delta\nu_h}{2\pi\left[(\nu - \nu_0)^2 + (\Delta\nu_h/2)^2\right]} \tag{23}$$

is the homogeneous line shape factor, $\Gamma_D(\nu - \nu_0)$ is the (time dependent) inhomogeneous lineshape function, and $q_{21} = \frac{7}{12}\frac{A_{21}\lambda_0^2}{8\pi}$ with A_{21} being the Einstein coefficient. The time τ_b describes the relaxation of the occupation numbers towards an equilibrium value $\bar{N}_i^{eq} = \Gamma_D^{eq}(\nu - \nu_0)\, N_i$, where $\Gamma_D^{eq}(\nu - \nu_0, t)$ is determined by the instantaneous Doppler width, which relaxes with time constant τ_v.

The corresponding rate equation for the photon flux F at $\nu = \nu_0$ reads:

$$\frac{d}{dt}F = -cq_{21}F\int_0^\infty \Gamma_h(\nu - \nu_0)\left[\bar{N}_2(\nu) - \frac{g_2}{g_1}\bar{N}_1(\nu)\right] d\nu - \frac{1}{\tau_r}F \tag{24}$$

where τ_r is the cavity life time.

The above rate equations do not contain all the necessary information to fully describe the relaxation of the velocity distribution. They do have the relaxation of the angular anisotropy with time constant τ_b, as detailed in Eq. (10). A separate procedure was utilized in the computer model to implement the affect of cooling on the occupation numbers as described at the end of the previous section. Every time step, the horizontal scale of $\Gamma_D(\nu - \nu_0, t)$ was changed according to Eqs. (14) and (15).
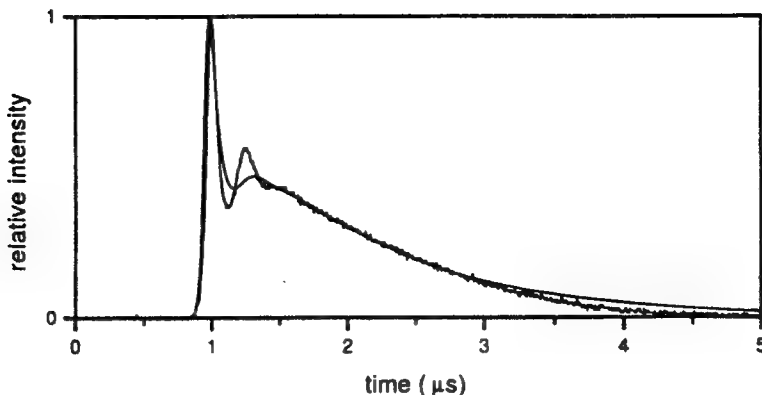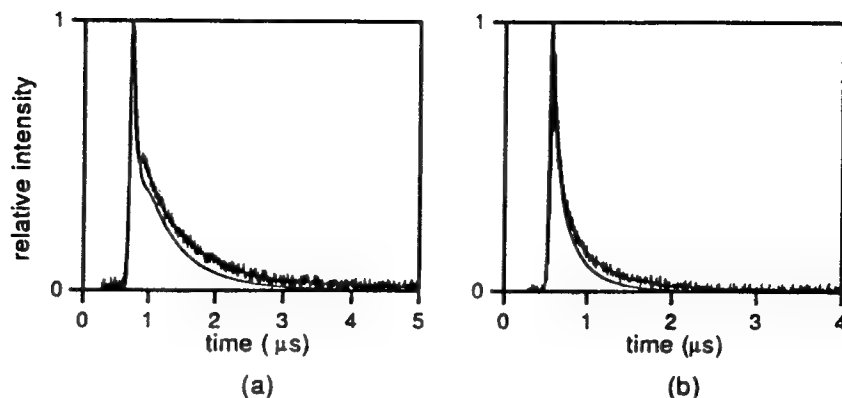
Figure 10: Experimental and theoretical pulse shapes for 0.8 torr CF_3I. The time constants used were $\tau_v = 650$ ns and $\tau_b = 120$ ns

The pumped region of the gain cell was a column 4 mm by 10 mm. Within this region approximately 15% of the CF_3I was dissociated into $CF_3 + I^*$. This column was surrounded by a bath of undissociated CF_3I. A diffusion equation calculation was done in order to see how long it would take the unexcited CF_3I to diffuse into the center region of the cell that comprised the mode volume of the iodine laser. We estimated it would take approximately 2.5 μs after the arrival of the pump pulse for the cold CF_3I to diffuse into the center region. This would have an effect on the tail of the pulse only (see Fig. 3), and therefore, diffusion was neglected in later calculations.

The initial values for the occupation numbers were set using Eq. (7). A value of $\beta = 1.4$ was assumed. The mean translational energy of the fragments of CF_3I after dissociation at 248 nm has been measured to be 10200 cm^{-1} for the $CF_3 + I^*$ channel [14]. This gives an initial Doppler width of 950 MHz for I^*. With a large bath of buffer gas, the Doppler width would relax to its room temperature value of 250 MHz. However, under our experimental conditions, we did not have a bath at room temperature. Instead, the excess energy in the dissociating photons is redistributed among the species in the gain cell according to their specific heats. Thus, with 15% of the 0.8 torr CF_3I being dissociated, the final Doppler width of the I^* was 330 MHz. With 2.4 torr of buffer gas in the cell, the final Doppler width was 290 MHz. The linewidth of the homogeneous transition $\Delta\nu_h$ used in Eq. (23) was calculated using published pressure broadening coefficients for the species present in the gain cell [15].

4   Comparison of Experiment and Theory

The rate equation model was used first to fit the experimental pulse shapes obtained with only CF_3I in the cell. The experimental and theoretical pulse shapes are shown in Fig. 10. The time constants used in

35-15

Figure 11: Experimental and theoretical pulse shapes for 0.8 torr CF_3I plus (a) 0.8 torr Xe, (b) 2.4 torr Xe. The contribution to the relaxation times $\tau_{v,b}$ due to Xe was decreased by a factor of three from (a) to (b)

the rate equation model were $\tau_v = 650$ ns and $\tau_b = 120$ ns.

The relaxation times τ_v and τ_b can be expressed as rate constants k_v and k_b respectively. In other words, the contribution to τ_v and τ_b is inversely proportional to the partial pressure, and the relaxation times can be written as

$$\frac{1}{\tau_v} = \sum_i \frac{1}{\tau_{v,i}} \tag{25}$$

with a similar equation for τ_b. The summation is over the various species present in the gain cell, i.e. CF_3I, CF_3, I, and the buffer gas.

Next, the data for 0.8 torr CF_3I plus the various buffer gas was fit. The contribution of the buffer gas to the relaxation times can be obtained using Eq. (25) and the relaxation times from the fit for pure CF_3I in the cell. For every buffer gas used, once the data for 0.8 torr CF_3I plus 0.8 torr buffer gas had been fit, the data with higher partial pressure of buffer gas was simulated simply by keeping the contributions to τ_v and τ_b inversely proportional to the buffer gas partial pressure. A typical fit for Xe as the buffer gas is shown in Fig. 11. Because the effects of Kr and Xe were the same to within experimental error (see Fig. 6) no effort was made to fit the Kr data separately. Table 1 contains a summary of the rate constants obtained for the speed and anisotropy relaxations and the ratio τ_v/τ_b for all buffer gases used.

The difference between the noble gases and the polyatomic molecules as buffer gas appears in the ratio τ_v and τ_b. Although N_2's atomic mass is only slightly larger than Ne's, the ratio of τ_v to τ_b is much greater, whereas the ratio for CO_2 is the smallest measured. This suggests that the hard sphere collision theory no longer applies for these polyatomic molecules.

Earlier measurements of He used as a buffer gas with C_3F_7I estimated relaxation times for He to be approximately 10 times as long as for C_3F_7I [2]. If we use this estimate along with the relaxation times

Buffer Gas	$k_v * 10^{11}$ cm^{3}mol^{-1}s^{-1}	$k_b * 10^{11}$ cm^{3}mol^{-1}s^{-1}	τ_v/τ_b
He	—	—	—
Ne	1.1	3.0	2.7
Kr	5.2	11	2.1
Xe	5.2	11	2.1
N$_2$	2.1	11	4.0
CO$_2$	3.7	6.5	1.7

Table 1: Rate constants obtained from the rate equation model
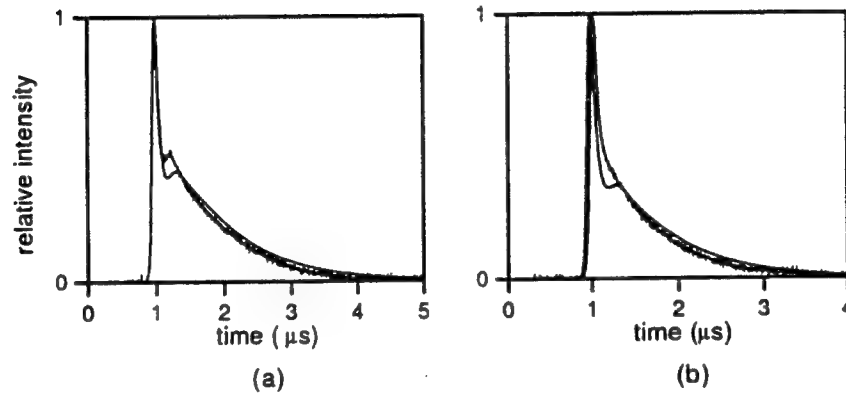

(a) (b)

Figure 12: Experimental and theoretical pulse shapes for 0.8 torr CF$_3$I plus (a) 0.8 torr He, (b) 2.4 torr He. τ_v and τ_b were not changed from (a) to (b). Only $\Delta\nu_D$ and $\Delta\nu_h$ were changed.

measured for CF$_3$I, and assume the relaxation times for C$_3$F$_7$I and CF$_3$I to be comparable, the change in the total relaxation times is too small to account for the measured change in the pulse shape as He was added to the gain cell. In fact, we were able to fit the measured pulse shapes by changing the homogeneous line width and final inhomogeneous linewidth only, without changing τ_v or τ_b (see Fig. 12). Thus, we concluded that the affect of He on the relaxation times was too small to be measured at our working pressures.

5   Conclusions

The pulsed output of an iodine photodissociation laser operating at a low pressure of the active medium has distinct features which are not explained in a simple rate equation model. These features include a long delay between dissociation of the CF$_3$I and onset of lasing, a long pulse duration as compared to the

cavity life time, and a distinct second peak in the iodine pulse. Furthermore, these features are dependent on the amount of buffer gas that is introduced into the cell, although the number of active iodine atoms is kept constant. As more buffer gas is used, the build up time and pulse duration decreases, and the second peak becomes a shoulder and finally disappears. We were able to model all of these phenomena using a comprehensive rate equation model that included the relaxation of the speed and the angular anisotropy in the velocity distribution and relaxation between the $^2P_{1/2}$ hyperfine levels. This model neglected effects due to diffusion and the existence of all but one (the 3-4) of six transitions in atomic iodine. Although the rate equation model was able to fit the build up times, pulse duration, and decay in the tail of the pulse strikingly well for all buffer gases, reproducing the exact shape of the second peak proved to be difficult. There are several possible explanations for this.

From the beginning, the separability of the velocity distribution was only an approximation, and one of the effects that tends to reduce the separability over time is collisions; the more massive the collision partner, the greater the effect [13]. Because rapid changes in $I(\theta, \phi, t)$ are being affected in the area of the second peak in the pulse, we expect the degree of separability of $F(\vec{v}, t)$ plays an important role in the dynamics of the second peak.

Furthermore, during the modeling of the data, it was found that the homogeneous linewidth was a critical factor in determining the shape of tail of the pulse, because it determines the width of the hole that is burned in the Doppler profile, and thus, how that hole is filled. In the model, it was assumed that the homogeneous linewidth was a constant. However, this is not the case. In the early time after dissociation, before the iodine atoms have had a chance to undergo many collisions, the homogeneous linewidth is poorly defined. And in later times, as the gas comes into thermal equilibrium and diffusion begins to play a more important role, the homogeneous linewidth must certainly be time dependent. Further experiments are required to determine the dynamics of the homogeneous linewidth.

Although we could not obtain relaxation times for pure CF_3I, since there was always a significant amount of CF_3I plus dissociation products in the active volume of the laser, we were able to obtain relaxation times for a number of buffer gases. In addition, we were able to successfully explain in detail the dynamics of the laser pulses as a function of buffer gas pressure. Our measurements confirmed that the heavier buffer gases were more effective in relaxing the velocity than in relaxing the angular anisotropy. In addition, the polyatomic molecules showed different ratios for τ_v and τ_b than the monatomic Noble gases.

References

[1] B. Shizgal and R. Blackmore. *Chemical Physics*, 77:417, 1983.

[2] J. I. Cline, C. A. Taatjes, and S. R. Leone. *J. Chem. Phys.*, 93:6543, 1990.

[3] R. Schmiedl, H. Dugon, W. Meier, and K. H. Welge. *Z. Phys. A*, 304:137, 1982.

[4] J. Nicholson, W. Rudolph, J. McIver, R. Tate, and G. Hager. *Applied Physics Letters*, 67:64, 1995.

[5] G.Brederlow, E.Fill, and K.J.Witte. *The high-power iodine laser.* Springer, 1983.

[6] H.J.Baker and T.A.King. *J.Phys.D:Appl.Phys.*, 8:L31, 1975.

[7] Heon Hwang, Min Hee Lee, and Sang Soo Lee. *IEEE J. of Quantum Electron.*, QE-18:148, 1982.

[8] V. A. Alekseev, T. L. Andreeva, V. N. Volkov, and E. A. Yukov. *Soviet Physics JETP*, 36:238, 1973.

[9] M.A.Kelly, J.K.McIver, R.F.Shea, and G.D.Hager. *IEEE J. Quantum Electron.*, QE-27:263, 1991.

[10] W. Thieme and E. Fill. *Optics Communications*, 36:361, 1981.

[11] R. N. Zare. *Molecular Photocemistry*, 4:1, 1972.

[12] G. N. A. Van Veen, T. Baller, A. E. De Vries, and M. Shapiro. *Chemical Physics*, 93:277, 1985.

[13] C. A. Taatjes, J. I. Cline, and S. R. Leone. *J. Chem. Phys.*, 93:6554, 1990.

[14] Peter Felder. *Chemical Physics*, 143:141, 1990.

[15] H. J. Baker and T. A. King. *J. Phys. D: Appl. Phys.*, 9:2433, 1976.

STEREOGRAPHIC IMAGING OF IRREGULARITIES
ASSOCIATED WITH HEATED REGIONS IN THE IONOSPHERE

Gary S. Sales and Arjang Noushin

Center for Atmospheric Research
University of Massachusetts Lowell
Lowell, MA.
USA

STEREOGRAPHIC IMAGING OF IRREGULARITIES
ASSOCIATED WITH HEATED REGIONS IN THE IONOSPHERE

Gary S. Sales and Arjang Noushin

Center for Atmospheric Research
University of Massachusetts Lowell
Lowell, MA.
USA

Abstract

The goal of this investigation is to demonstrate a technique for three dimensional imaging of irregular electron density structures in the F-region of the ionosphere, such as polar patches, high power HF ionospheric modification regions and equatorial depletion bands. Through the use of multiple HF sounders, each capable of locating ionospheric irregularities or sources using the standard Digisonde Doppler interferometry technique, it is demonstrated that the array of sounders all "view" the same irregular regions of the ionosphere and in the process measure the velocity components in the direction of each sounder. Combining these data from many sources leads to a model of the velocity distribution within the disturbed region.

STEREOGRAPHIC IMAGING OF IRREGULARITIES
ASSOCIATED WITH HEATED REGIONS IN THE IONOSPHERE

Gary S. Sales and Arjang Noushin

Introduction

Digisondes have been used over the past many years to investigate the structure of ionospheric phenomena from the Arctic to the equator. Besides making vertical ionograms for characterizing the electron density profile at the site, the Digisonde, operating in the "drift" mode in-between ionograms, uses the technique of Doppler interferometry, developed at the Center for Atmospheric Research, to identify and locate individual reflecting sources in the ionosphere. The scale of the observed sources is typically of the order of a Fresnel radius, i.e., 6 km. Normal operation involves a single sounder operating in the vicinity of structured ionospheric irregularities such as polar patches, ionospheric modification regions and equatorial depletion bands. The effort, here, is concentrated on probing artificially heated regions of the F-region of the ionosphere where electron density irregularities in the plasma are generated by processes that transfer energy deposited by the high power heater transmitter into the irregular region. The techniques developed here are equally applicable to the other types of disturbed regions.

This investigation is aimed at developing a stereographic imaging technique that makes it possible to determine, besides the location of the irregular structures, the distribution of velocities associated with the ionospheric irregularities formed during the heating process. The simulations carried out here involves two sounders separated spatially by about 90° in azimuth and each some 50 to 100 km from the heater site. The difficulty with using the standard Digisonde drift analysis technique for these disturbed regions is that the sources produced by the heater are more confined spatially than would be the case for naturally occurring ionospheric irregularities associated, for example, with spread-F.

HF Digisondes, both ground based and airborne, have supported a series of ionospheric modification experiments that have been carried out over the past several years; particularly at the Arecibo heating facility in Puerto Rico. These measurements consisted of ionograms and drift mode operation, providing the basic background ionospheric conditions and some information about the existence of irregularities in electron density that form shortly after the high power pump wave is turned on. The spatial extent of the pump wave in the F-region of the ionosphere is determined, basically by the "heater" antenna pattern. Here it is assumed that the heater illuminates a 100 km x 50 km area in the F-region at an altitude of 250 km; the long axis is in the N-S direction.

The energy in the pump wave is coupled to the ionospheric plasma in the altitude regime in the F-region where the pump frequency matches the plasma frequency (X=1) and the upper hybrid frequency (X=1-Y). Here, the coupled energy stimulates the growth of ionospheric irregularities with scales from the order of kilometers down to meters. Depending on the sensor, the measurements are particularly sensitive to a narrow range of scale sizes. For example, back scatter systems are, in general, affected by the scales corresponding to $\lambda/2$ while forward scatter systems interact with the largest scales in the scattering region. Motions of the these irregularities result in a Doppler frequency spreading of the scattered signal. The spreading also depends on the mode of operation of the particular sensor. In general, forward scatter modes suffer little Doppler spreading relative to back scatter soundings.

Currently, plans are being made to experimentally monitor the formation and growth of these electron density irregularities at the Arecibo facility in Puerto Rico and at the HAARP facility in Gulkana, Alaska. The study presented here is aimed at simulating HF sounding and HF/VHF radar probe experiments to better plan the operation and better understand the results. Using a simple model of the heated region of the ionosphere; i.e., assuming it contains a full spectrum of irregularities and together with a three dimensional ray tracing code, calculations are carried out to simulate both vertical soundings as well as oblique HF and VHF (50 MHz) radar probing.

For the situation when the irregularities are produced by high power heater waves, the velocities are expected to be relatively random and spatially restricted. The proposed stereographic technique allows the "same" irregularities to be observed from different directions and a resultant horizontal velocity component uniquely determined for each source. Then the distribution of sources leads to the characterization of the Doppler spectrum spreading of the scattered signal.

Simulation of the Stereographic Imaging Technique

Vertical Sounding

The technique for simulating scattering from small scale field aligned electron density irregularities in the F-region involves finding the locations where the ray tracing code indicates that the rays have achieved orthogonality with the earth's magnetic field using a model of the local ionosphere that includes only large $(L \gg \lambda)$ structures. It is impossible to include all the scales of irregularities involved in scattering or reflection as this would overload the computer memory and in some cases, be inconsistent with the assumptions of the geometrical optics. However, with the assumption that all scales of irregularities are aligned with the earth's magnetic field, finding the ray orthogonality with the magnetic field is sufficient.

The results of the simulations carried out here for the Arecibo heater location apply equally well to the HAARP system at Gulkana, Alaska since at the higher latitude of Gulkana the magnetic field has a steeper inclination than at Arecibo and achieving ray orthogonality with the earth's field is even easier. For Arecibo the magnetic dip angle is approximately 45° and the magnetic declination is small, about 1° or 2°. As stated earlier, the heated region is defined in latitude and longitude by the parameters of the heater antenna array and the altitude extent by the heater frequency and the ambient electron density profile. The altitude extent is usually limited on the bottom side by the upper hybrid frequency and in the model selected here for study, based on PIM, the lowest altitude for forming ionospheric irregularities is about 235 km. The calculated vertical ionogram and the true height profile, with a foF=3.67 MHz at an altitude of 295 km are shown in Figure 1. The altitude of the upper hybrid and plasma frequency resonances are indicated in the figure for a heater frequency of 3.2 MHz.

For the Arecibo heater experiments, to demonstrate the stereographic imaging technique, two sounder locations were selected. The first uses the existing DISS system at Ramey (18.5°N, 292.7°E) which lies some 50 km west of Arecibo. The second site chosen for a new DPS is at Ponce (18.0°N, 293.3°E) lying about 50 km south of Arecibo. No attempt is made at this point to determine an optimum distance from the heated region for locating the imaging sounders. As is shown later, this particular choice works well, though distances out to 100 km or even further may be as good. This is a subject for a future expanded investigation. In the case of Arecibo, on the island of Puerto Rico, 50 km is about as far away as it is possible to move and remain on the island. For the HAARP program, other choices are possible and should be investigated.

For these simulations, the heated region was defined as centered over Arecibo and extending ±1° of latitude north and south and ±0.5° of longitude east and west of the site (at 3 dB down from the peak of the heater beam). The two to one ratio is determined by the Arecibo high power transmitting antenna array. Figure 2 is a map of Puerto Rico showing the two sounder sites and the heated region projected radially downward onto the ground. Ramey sits just on the western edge of the heated region, while Ponce sits to the south, somewhat within the heated region. The lower height of the heated region is set at 235 km, corresponding to the altitude of the upper hybrid frequency.

Ray tracings were carried out from both sites over the frequency range of 2 to 6 MHz and at all azimuths and elevation angles that lead to rays passing through the defined boundaries of the heated region. For both Ramey and Ponce, elevation angles from 90° down to 45° was sufficient to capture all the relevant orthogonalities. These data can be displayed in several ways that convey different aspects of the simulation process. The first and probably easiest to understand is to plot the location of the
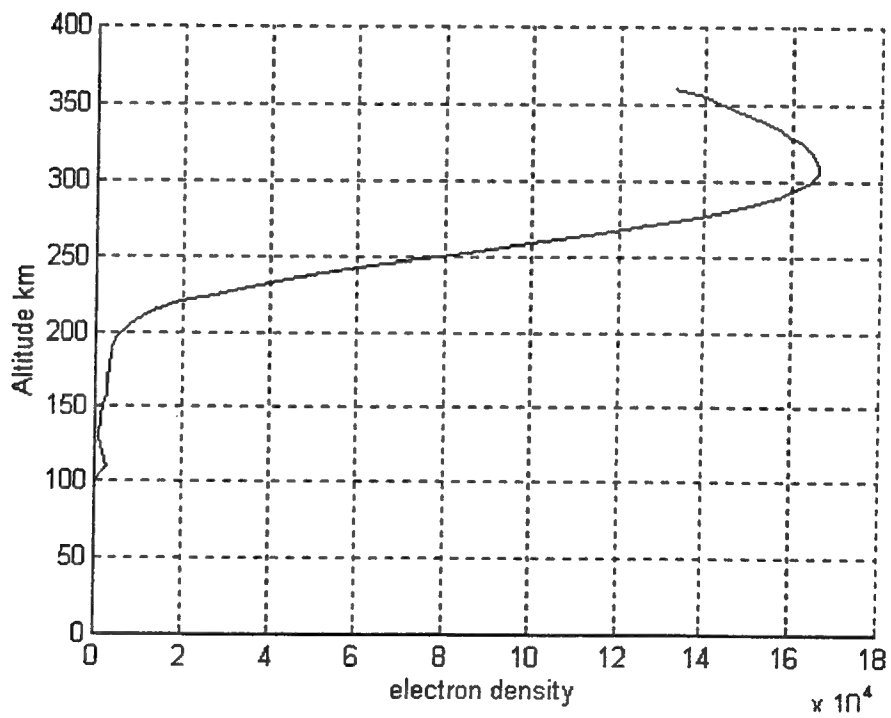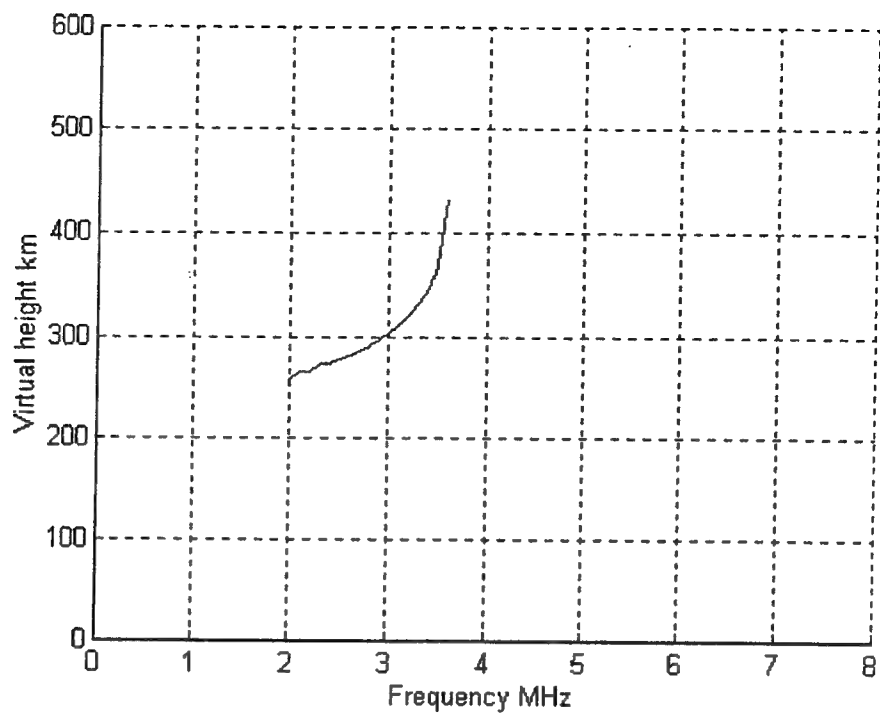
Figure 1. Vertical electron density profile and calculated vertical incidence ionogram for Arecibo, Puerto Rico at 02 UT in winter (SSN = 50).
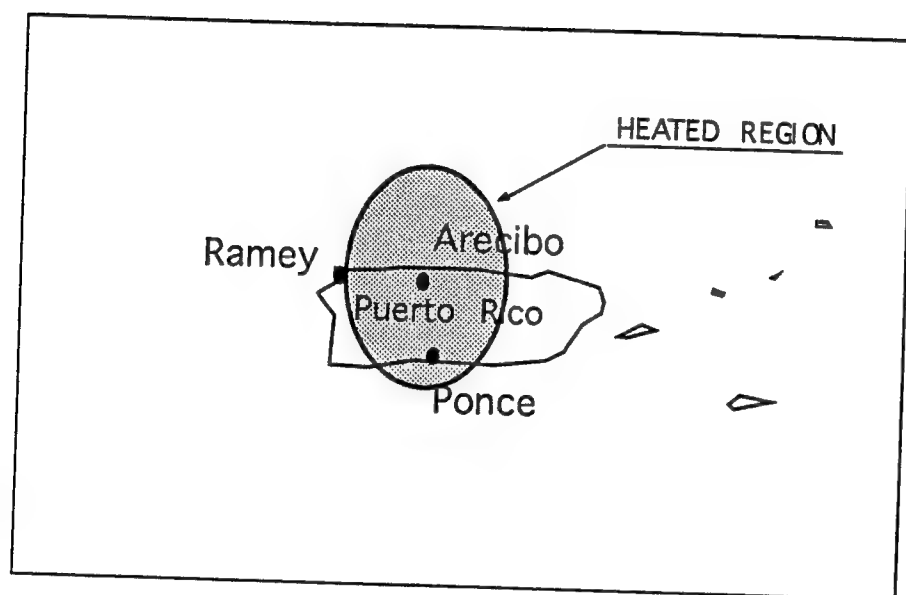
Figure 2. Map of Puerto Rico showing the selected sounder sites and the projected heated region.

orthogonalities in a three dimensional latitude, longitude, altitude grid. Figure 3a, for the Ramey site, shows that this sounder can detect field aligned irregularities anywhere within the heated region from the lowest altitude of 235 km up to about 270 km, some 20 km below the peak of the layer. Figure 3b, which is a N-S cut through the heated region, shows that for frequencies from 2.5 MHz to 3.5 MHz, i.e., below foF, complete coverage over the heated region is achieved. For the selected time (02 UT), season (winter) and sunspot number (50) the F-layer critical frequency is 3.67 MHz and higher frequencies tend to penetrate the layer. For the lowest frequency, 2 MHz, the orthogonalities all lie below the lowest altitude for irregularity formation, in this case 235 km and no scatter is indicated.

For frequencies above foF, orthogonalities are found only to the north of the heater site, of course within the heated region. For these penetrating rays, orthogonalities are found both below and above the peak height of the layer at 295 km. Combining measurements from frequencies below and above the layer critical frequency offers the possibility to investigate the growth of the irregularities through the peak of the layer. This effect is enhanced if the heater beam is steered somewhat north of the overhead position as can be seen in Figure 3b.

Figure 4a and 4b are the same plots for the Ponce sounder. Here, the situation is very similar, with full coverage of the heated region for frequencies below foF and some limited geographical coverage for frequencies above foF up to 6.0 MHz. The small increase in the top frequency as seen at Ponce results from the greater N-S extent of the heated region, i.e., the greater distance to the furthest point makes it possible to use a somewhat higher frequency.

An important question concerning stereographic imaging arises at this point. Can the same irregularity region be observed by both sounders? The relatively uniform distribution of orthogonality points in Figures 3 and 4 makes it easy to conclude that this is, most likely, the case. To demonstrate this, an orthogonality point was selected at random from the Ponce ray tracing and a search was made to find the ray from the Ramey site that has its orthogonality at the same point in space. This was easily accomplished by trial and error and the two rays are shown in Figure 5. Extending this reasoning, it is expected that the actual Digisonde source location data from the two sites can be combined to locate the scattering irregularities and to determine the horizontal component of the drift velocity associated with each area. Then all of these data can then be combined to generate a three dimensional image of the disturbed region, including the uniform and random motions of the irregularities.
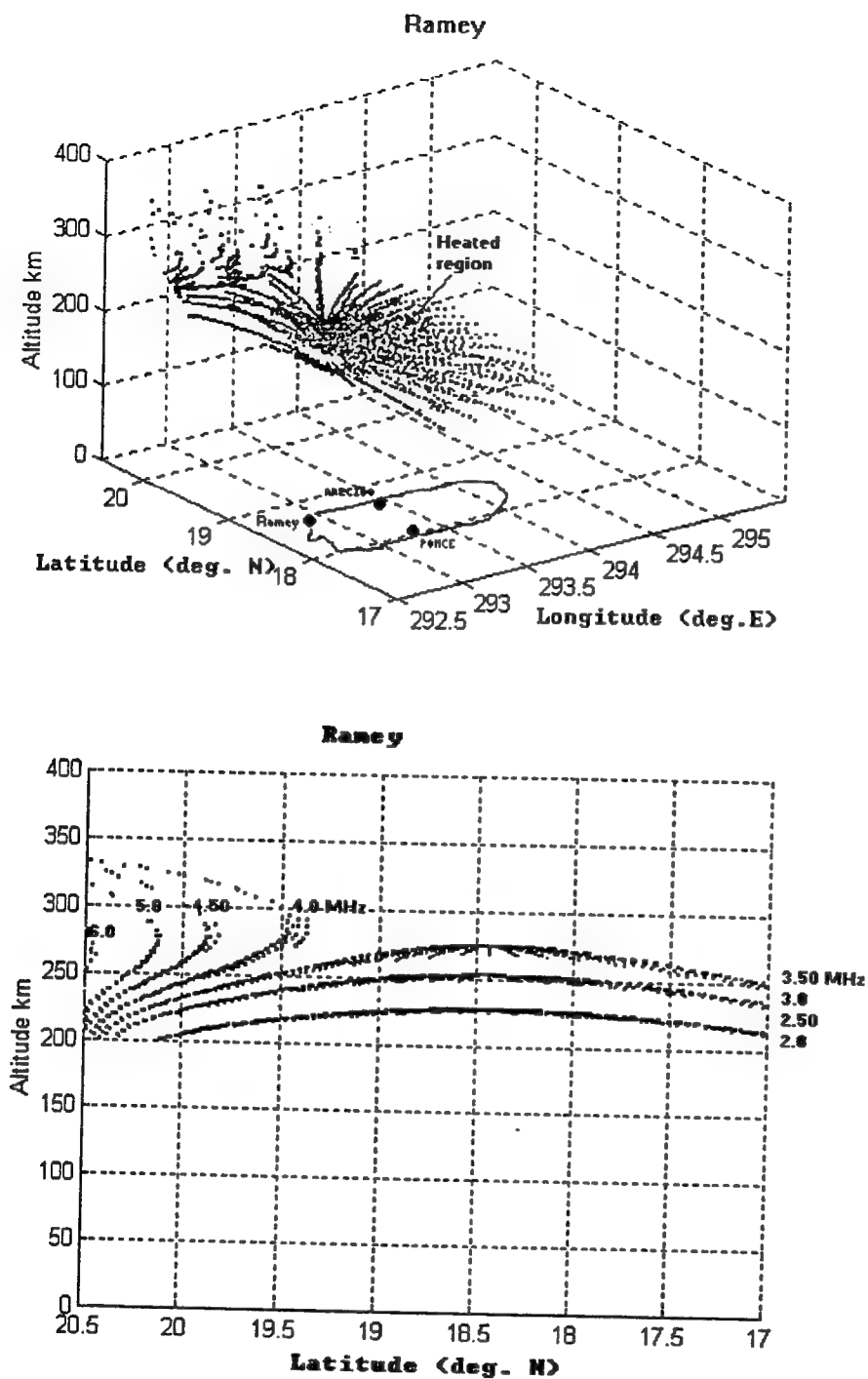
Figure 3. a) Orthogonal source locations for the Ramey sounder in the heated region as a function of frequency; latitude/longitude/altitude grid.

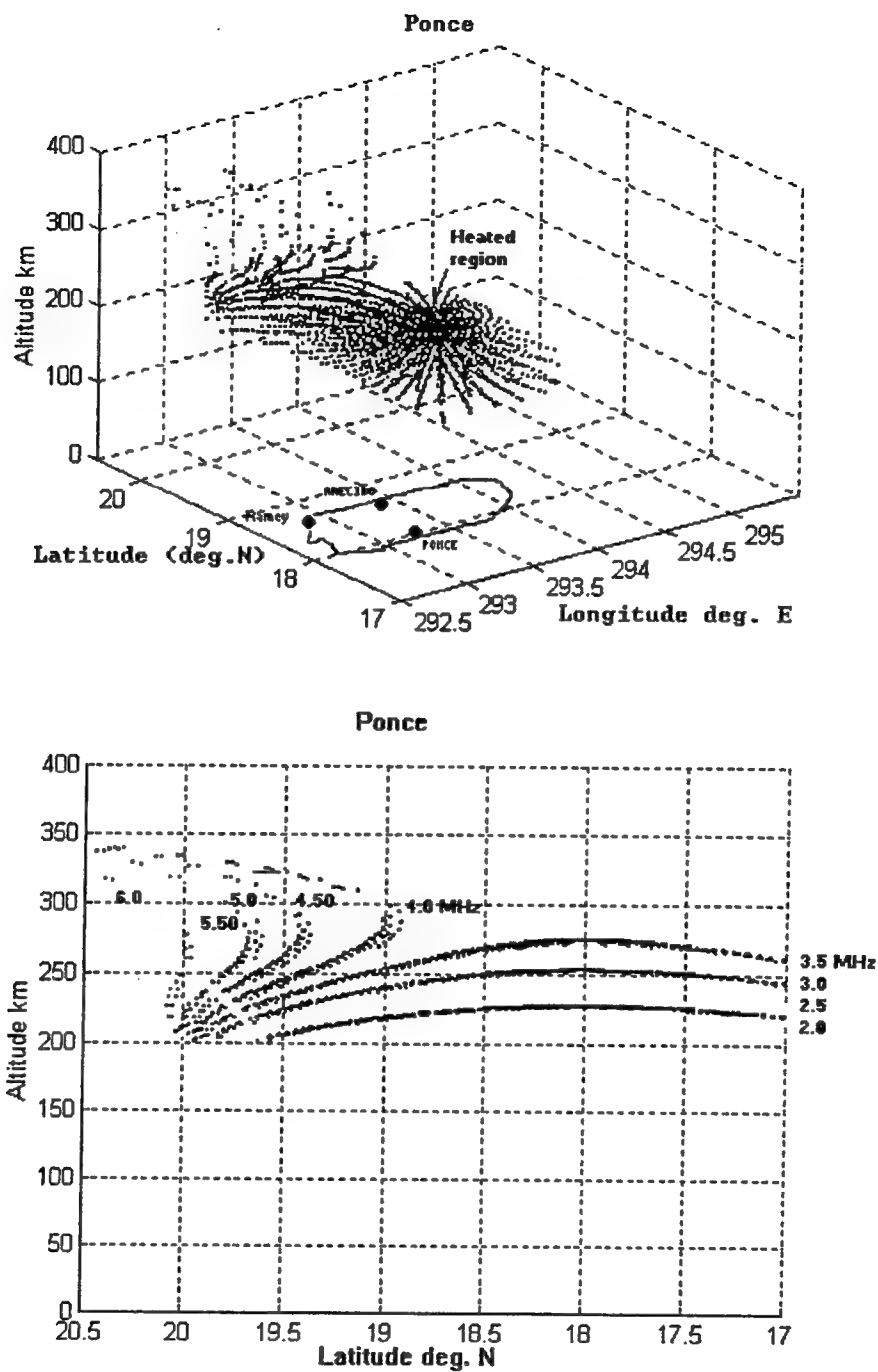　　　　b) North-south cut through the orthogonal points in the heated region.

Figure 4. a) Orthogonal source locations for the Ponce sounder in the heated region as a function of frequency; latitude/longitude/altitude grid.

b) North-south cut through the orthogonal points in the heated region.
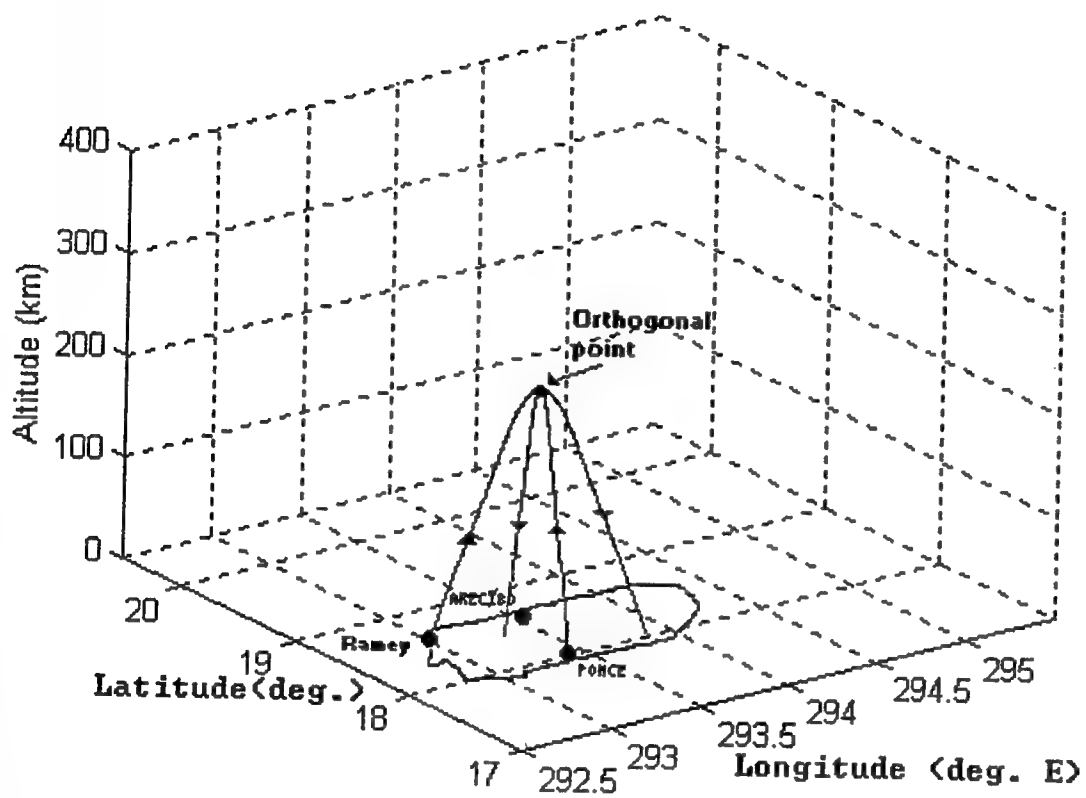
Figure 5. Ray tracing from Ponce and Ramey having the same orthogonal point in the heated region.

The second display technique involves using the simulation data to generate a vertical ionogram showing the background overhead trace as well as spread-F traces that result from signals scattered/reflected from the heated region. Here, besides retaining the altitude of the orthogonality point, the group delay of each orthogonality was recorded and used at each frequency (2 to 6 MHz in 0.5 MHz steps) to generate a spread trace on the ionogram. The two simulated ionograms, for Ramey and Ponce, are shown in Figure 6. The results at the two sites are similar, differing primarily in the extent of the spread frequency trace above foF. This difference comes about because for Ponce the northern edge of the heated region is further away from the sounder than is the eastern edge from Ramey.

The important result is that the scatter trace from the heated region extends above foF by more than 1MHz. and the shape of the spread traces makes it easily distinguishable from the overhead trace even though the sounders are only 50 km from the heater site.

These simulations differ from those in the earlier work of Buchau and McNamara where they also presented actual ionograms made aboard the US Air Force Ionospheric Observatory Aircraft flying in the vicinity of Arecibo. Most of the ionograms presented there where for ranges considerably greater than those used in this simulation, typically several hundred kilometers away from the heater site. The simulations presented in this paper also differ from the Buchau work in that for their simulations they ray traced only to the center point of the heated region and they did not consider the extended area of ionospheric irregularities. This makes their results look different in that their signal returns from the heated region appear as a narrow trace rather than the spread trace calculated here.

Oblique Sounding

Oblique sounding offers the possibility to achieve orthogonality with the magnetic field above Arecibo at higher frequencies that suffer little refraction while passing through the background ionosphere. Back scatter at these higher frequencies makes it possible to detect smaller scales of irregularities than from any site on the island of Puerto Rico. For example, with a 50 MHz back scatter radar it is possible to monitor the development of the 3 m scale of irregularities. However, the radar site must lie on the intersection of the plane that is perpendicular to the earth's magnetic field in the F-region of the ionosphere with the earth's surface.
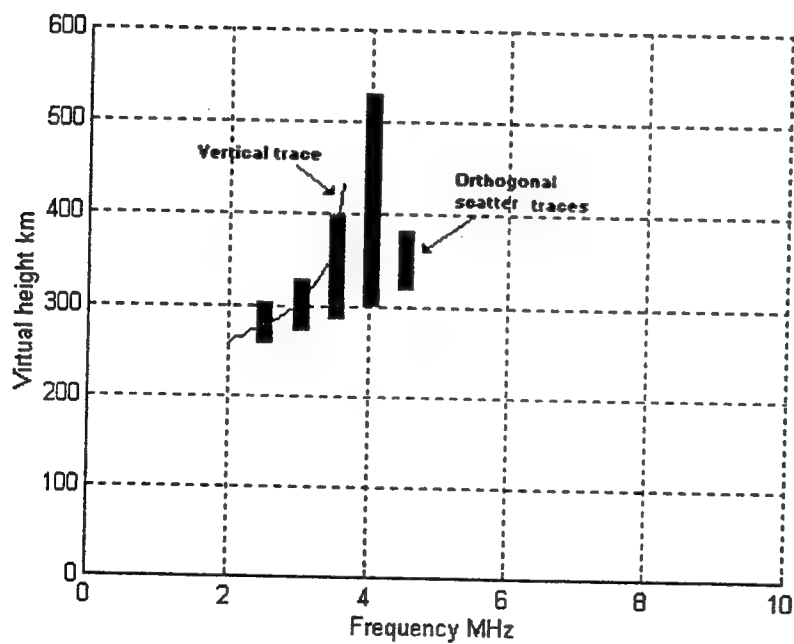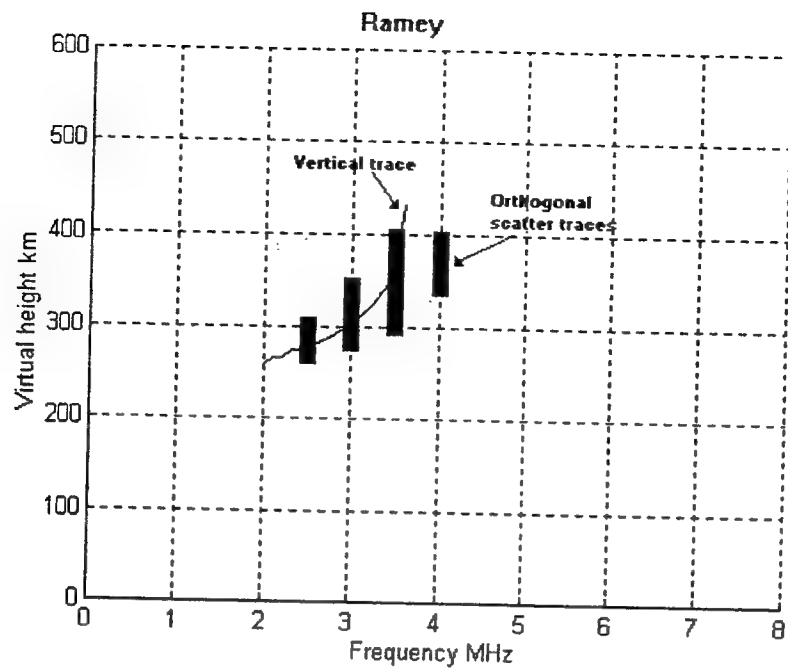
Figure 6. Simulated vertical ionograms from Ramey and Ponce showing spread returns from the heated region superimposed on the background vertical trace.

Simple geometrical calculations show that for Arecibo, the plane intersects the earth along a circular cut (Figure 7) beginning magnetically south of Arecibo and extending in an almost east and west direction from that point. The sparcity of islands along the circular path leads to the selection of the island of Guadeloupe (16.3°N, 298.5°E), some 650 km from Arecibo, as one of the few acceptable sites. From Guadeloupe, a straight line ray intersects the magnetic field above Arecibo at an altitude of 250 km at right angles.

Ray tracings at frequencies from 8 to 50 MHz were carried out from a site on Guadeloupe in the general direction of Arecibo, azimuthally covering the entire heated region and for elevation angles from 12° to 20°. Figure 8a shows the location of the orthogonalities in the latitude, longitude, altitude grid for all frequencies again indicating that good coverage is possible from a remote site, though not as complete as with the near-vertical soundings. No orthogonalities are found in the northern half of the heated region. At the highest frequency where refraction is relatively small, the orthogonalities form a surface corresponding to the plane surface that was used to compute the locus on the ground. This surface runs through the center of the heated region. At all lower frequencies these surfaces are no longer plane and they lie to the south of the high frequency plane. At these lower frequencies, ionospheric refraction is important and the orthogonal surfaces are more complicated and reveal that each frequency, by itself "sees" only a small portion of the heated region. Together, they provide good coverage only over the southern half of the heated area.

Although it appears that back scatter experiments from the island of Guadeloupe can work well at most frequencies (8 to 50 MHz), forward or side scatter experiments are another matter. Most of the high frequencies that reach the heated area, some 650 km away, at altitudes above 235 km (lower frequencies have their orthogonalities below the heated area) have no reflection and ultimately penetrate the ionosphere. Under these circumstances where the ray is perpendicular to the earth's magnetic field, the scattering cone has opened into a plane and the strongly scattered signals fall along the orthogonal locus found earlier in the process of locating the radar site. As discussed earlier, there are no other sites along this locus that could serve as a receiving site for the strong scattered signal from the heated region above Arecibo.

This general approach of modeling and ray tracing can be used in the future to consider other sites where the incident ray is not orthogonal to the field aligned irregularities above Arecibo to find where the scattered signal does fall on an available receiving site.
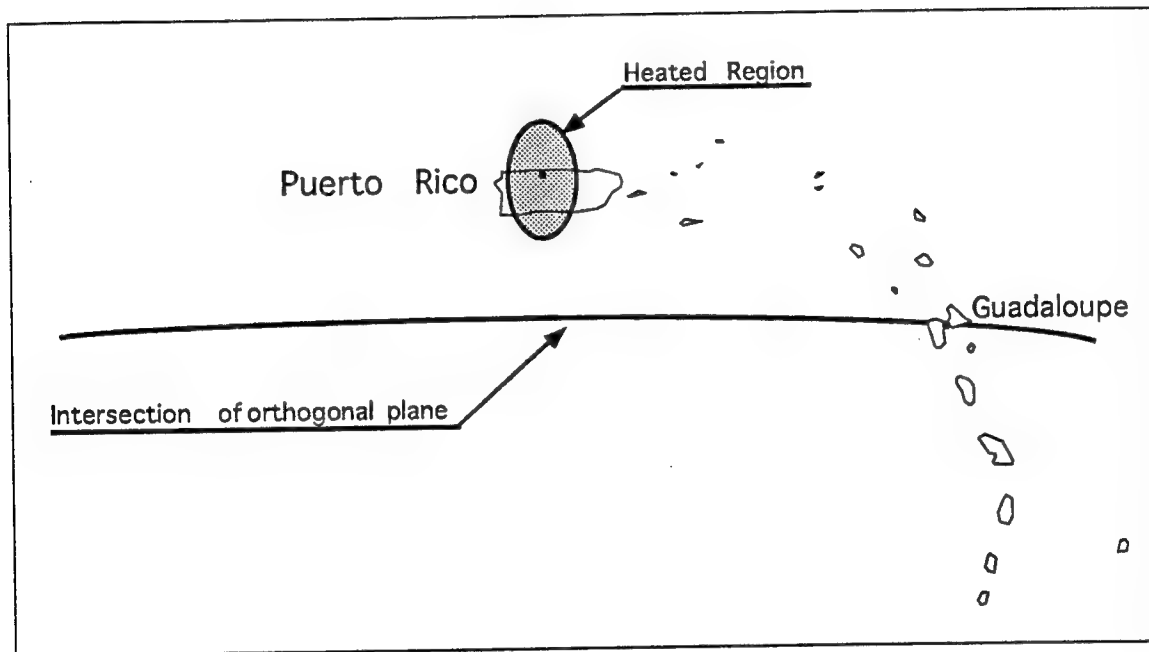
Figure 7. Intersection of the plane orthogonal with the earth's magnetic field over Arecibo with the surface of the earth.
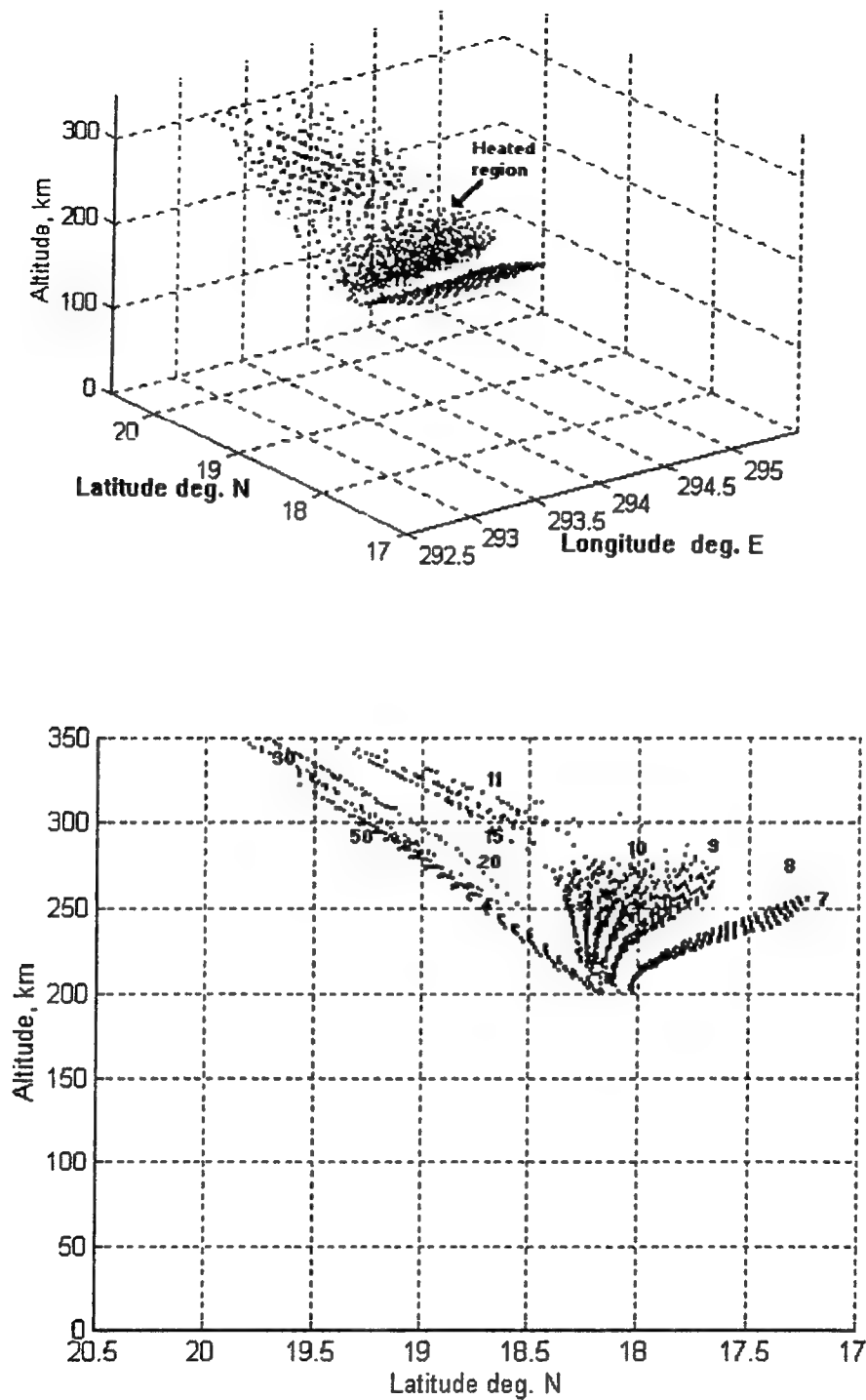
Figure 8. a) Orthogonal source locations for a Guadeloupe radar in the heated region as a function of
frequency; latitude/longitude/altitude grid.
        b) North-south cut through the orthogonal points in the heated region.

36-16

Conclusions

This stereographic imaging technique can serve as a means of describing the internal structure of ionospheric disturbances. With the understanding provided by these simulations, a clearer picture of the meaning of the radio sounding is possible. Ultimately a third sounder may be required to determine, unambiguously, the velocity vectors associated with each irregularity region within the structure. A better sounder configuration, useful for experiments at other facilities besides Arecibo, would be to locate two of the sounders some 100 to 150 km from the heater with a 90° bearing separation to determine the horizontal velocity components and then locate the third sounder, to measure the vertical component of the velocity, as close to the heater facility as is possible given the very high heater signal level.

Combining the results from the array of sounders yields a complete picture of the location and motions of the individual irregularity regions. Together, the sounder network provides a comprehensive characterization of the growth and decay of the irregularities in electron density in the disturbed region of the ionosphere.

The first opportunity to evaluate this technique will be in Puerto Rico in January, 1996. At that time only two sounders will be used, one at Ramey and the other at Ponce. Similar simulations are going to be carried out for the HAARP facility in Alaska.

SPECTRAL EFFECTS ON THE SIGNAL-TO-NOISE-RATIO OF SHORT EXPOSURE
PHOTON-LIMITED STELLAR DATA

Melani Shoemaker
Instructor
Department of Electrical Engineering

Seattle Pacific University
Seattle, Washington 98119

SPECTRAL EFFECTS ON THE SIGNAL-TO-NOISE-RATIO OF SHORT EXPOSURE PHOTON-LIMITED STELLAR DATA

Melani Shoemaker
Instructor
Department of Electrical Engineering
Seattle Pacific University

Abstract

Due to the desire for high resolution images of objects in space, several techniques are undergoing active research to combat the resolution limitations induced by atmospheric turbulence. Among these methods, speckle imaging techniques, initiated by Labeyrie, combine multiple short exposure images to reduce atmospheric effects. One significant measure of the success of this technique is the signal-to-noise ratio (SNR). Traditionally, efforts to increase this important quantity have been based on temporal issues. Our focus, on the other hand, is the investigation of the spectral bandwidth's impact on the SNR. This investigation involves comparison of theoretical expectations with computer simulated data and field data.

SPECTRAL EFFECTS ON THE SIGNAL-TO-NOISE-RATIO OF SHORT EXPOSURE PHOTON-LIMITED STELLAR DATA

Melani Shoemaker

1. Introduction

Speckle imaging methods allow retention of information at spatial frequencies up to the diffraction limit. A measure of the quality of an estimated image obtained by this method is the single frame power spectrum signal-to-noise ratio (SNR). The value of the SNR is directly proportional to the average number of photons per frame used in the speckle imaging calculations specified by Labeyrie. (2) The traditional method involves collecting light over a narrow spectral bandpass. Therefore, to increase the SNR, one increases the number of photons per frame by enlarging the frame time span. This SNR increase is linear with the number of photons until the SNR levels off when the photon count becomes large. When multiple frames are used, the data SNR becomes the single frame SNR multiplied by the square root of the number of frames. A fixed observation time and an elongated frame time span constrain the available number of frames and, therefore, the SNR. Recently, though, Fried and Hench of the Optical Sciences Company (tOSC) developed theory indicating that the number of photons per frame can be increased by collecting light over a broader range of wavelengths resulting in an increase of the average single frame SNR in the mid spatial frequency region. As the bandpass broadens to large ranges, the SNR will asymptote to a constant, but never decrease. (1) Thus, the SNR can be increased with more photons without the time span enlargement that limits the number of potential frames. This theory has yet to be verified experimentally; that validation was the goal of this summer's research.

In December 1994, stellar field data was collected through various spectral bandpass filters for the purpose of validating Fried and Hench's theory. To normalize the field data for comparison with the theory, a computer simulation was developed in the spring to model the effects of the atmosphere and telescope. The model incorporates atmospheric phase effects, photon noise, the telescope transfer function and spectral dependencies. In addition to its normalization utility, the computer simulation program is intended to enable research on the

effects of spectral bandpass on all parts of the speckle transfer function SNR, beyond just the mid spatial frequencies considered by tOSC. (1)

This report will summarize the data reduction process, tOSC's spectral SNR theory, and the simulation algorithm. We will present and compare the resulting power spectra and power spectrum SNRs, noting the discrepancies between the simulated and field data sets. Unfortunately, the match between the field data and the simulation is not yet sufficient for theory validation; however, we are able to make qualitative observations. Finally, we conclude with plans for future work.

2. Methodology

the field data: The field data used in this research was taken by Keith Hege of Steward Observatory using one mirror of the Multiple Mirror Telescope (MMT) at the University of Arizona. The primary mirror used in the collection was 1.8m in diameter and the secondary, 0.30m. Nine datasets were collected using the star Alpha Leo which is classified B7 with a surface temperature of 13000K. Two of the nine data sets employed a 30nm bandwidth, and the other seven sets had bandwidths of 40, 50, 70, 100, 140, 190 and 300nm. The maximum wavelength of each filter was 650nm. The detector employed was the Phillips Lab's photon-counting Multiple Anode Microchannel Array (MAMA). The MAMA camera provides temporal and spatial coordinates (t,x,y) for each event detected at rates up to 1MHz. This allows for data reduction at a user-selected time span per frame. We will select short exposures on the order of 10ms duration. Unfortunately, difficulties arose in the data recording instruments and the actual data recorded was only approximately 1 in 40 actual events. We assumed that this anomaly did not affect the photon statistics of the data set.

data reduction: The short time span of the field data allows us to use Labeyrie's discovery that high spatial frequency information is present in short-exposure atmospherically distorted images; however, the SNR at these frequencies will be low. (2) To boost the SNR, we calculated the power spectrum (the square of the modulus of the Fourier transform of the image) of each frame and averaged over many short exposure frames to raise the SNR. This rise will be a multiplicative factor of at least the square root of the number of frames used in the calculation. This power spectrum will be biased by the photon noise, which is equivalent to the number of events in the frame. After subtracting this bias from the power spectrum, running totals are

calculated of both the power spectrum and its square. These totals are manipulated in the same manner as our data sets from a year prior. "The mean power spectrum for the entire data set is the unbiased power spectrum running total divided by the number of frames in the data set. The power spectrum variance is also calculated as the square of the mean power spectrum subtracted from the running total of the square of the unbiased power spectra divided by the number of frames." (3) The average single frame power spectrum SNR can be formed, then, as the power spectrum mean divided by the square root of the variance. The power spectrum SNR is the focus of this research.

spectral bandpass/SNR theory: As stated earlier, it is well known that short-exposure snap-shots collected with a narrow spectral bandpass will retain high spatial frequency information. As the bandpass broadens, spectral decorrelation increases which would tend to lower the SNR. On the other hand, the broadening bandpass provides an increase in the number of photons/frame which counteracts the spectral decorrelation effects. Fried and Hench noted that the power spectrum SNR depends on the product of these two factors, namely the turbulence limited modulation transfer function in spatial frequency and the number of photo-events/frame. (1) Therefore, as long as this product increases, the useful information also increases regardless of the decorrelation effects. Their work showed mathematically that this product will never decrease with broadening spectral bandwidth, although it will asymptote to a constant. They theorized, therefore, that for a fixed number of frames and a fixed time per frame, enlarging the spectral bandwidth will never decrease the power spectrum SNR. (1)

It should be noted that this theory includes a few assumptions that must be remembered and incorporated in our simulation. First, their work does not account for temporal effects and thus implicitly assumes a very short time exposure for each frame. (1) Therefore, reduction of real data must use a short frame time span. Second, they make an approximation that assumes that the spectral brightness does not vary significantly over the spectral band involved. (1) We must consider this fact with respect to our field data. Third, the theory assumes that photon noise dominates which creates the requirement that the SNR at mid-range spatial frequencies must not be larger than approximately 0.1. (1) Finally, their theory is valid for starting at a particular center wavelength and expanding on either side of it, rather than the fixed upper wavelength of the field data. This difference should not affect the qualitative results. We anticipate that it

would induce a shift in the location of the knee between the linear rise of the SNR and the asymptotic approach.

the need for data normalization: Fried and Hench's theories are founded on the presence of increased photon numbers when the spectral bandwidth increases. (1) As described above, however, in collecting data one desires to obtain the highest photon rate allowed by the recording process for each data set in order to reduce noise effects. Consequently, each data set will have relatively the same photon rate regardless of the specific spectral filtering. Since this absence of increased photo-events was created in the data collection process, computer simulation will be used to appropriately modify the field data reduction results to allow comparison with the theory.

spectral effects on the transfer function: The current spectral SNR theories consider only the midrange spatial frequencies. (1) We intend to look at the entire transfer function to see spectral effects across the full frequency range. With the field data, we can compare results due to various time durations at the different spectral bandwidths. The simulation will allow us to explore various spectral implications.

To validate the spectral SNR theory, we would specifically compare the normalized SNR at a particular mid-range spatial frequency for each spectral bandwidth in our normalized field data. At the low bandwidths we should expect to see a linear rise in the SNR with possibly an asymptote to a constant at the higher bandwidths. The asymptote would only arise if the knee of the curve lies within the range of bandwidths present in the data.

modeling: In order to normalize the field data, we must first simulate it. To do this, we can think of the photon-limited stellar image obtained from a ground-based telescope as a true image (a point source) distorted by the atmosphere, filtered with a spectral bandpass filter, and modified by the telescope. The simulation models the atmosphere as a thin phase screen in the pupil plane of the telescope. The phase screen is created with Kolmogorov characteristics and Gaussian noise. Our simulation accounts for the wavelength dependence of the Kolmogorov filter, albeit with the approximation that r_o goes as λ rather than the true $\lambda^{6/5}$. The telescope pupil function accounts for the size of the MMT's primary and secondary mirrors. Temporal effects are not simulated. The simulation does incorporate the spectral variations in the irradiance of the object being imaged, the detector and the spectral filters. The object's spectral irradiance is modeled by the black body curve as estimated by Planck's law.
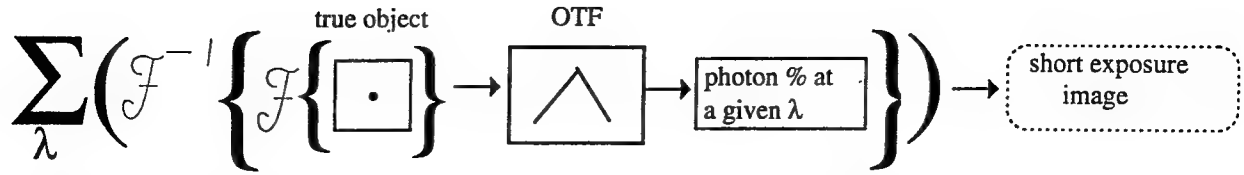
creating simulated data frames: The spectral band is broken into many narrow spectral regions. An image is created for each region, and a final image frame is the sum of these images. For each spectral range, the appropriate light intensity is estimated as a percentage of the total number of photons/frame (matched to the corresponding number in the field data). For each narrow region, a product is calculated of the object irradiance, the spectral filter transmission, and the MAMA camera quantum efficiency at the center wavelength of the region. The percentage of the total light is then this product divided by the sum of these products as calculated at each of the spectral regions in the entire band.

In creating an image at each spectral region, first the generalized pupil function is formed as the product of the phase screen and the telescope pupil function. Autocorrelating the generalized pupil function generates the optical transfer function (OTF) which is the appropriate transfer function since the object is self-illuminating with incoherent light (figure 1). After multiplying the Fourier transform of the true object by the photon percentage and the OTF, the Fourier transform of the created image is complete. The simulation algorithm calls for inverse Fourier transforming this result and then summing this narrow bandpass short-exposure image with those created at each of the other spectral regions (figure 2). When the spectral band has been covered completely, Poisson noise is applied to corrupt the short-exposure image to model the photon noise. Finally one frame of data is complete. Each of these frames should now resemble the field data. The simulated data set, accordingly, is reduced in the same manner as described earlier.
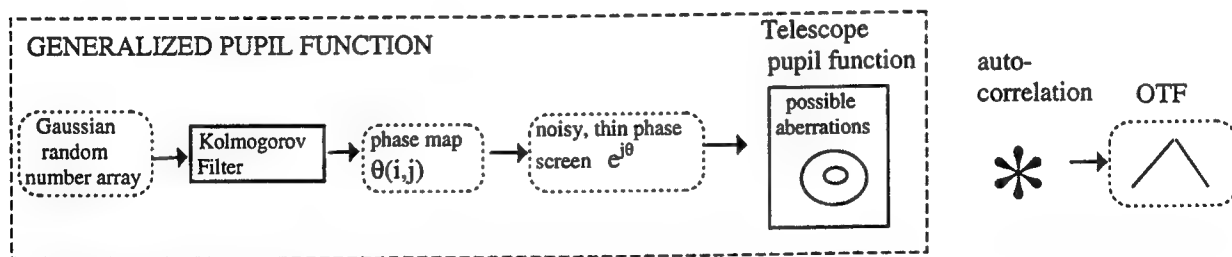
Figure 1 The process of modeling the atmosphere and telescope.
A wavelength dependent model of the atmosphere and telescope system for incoherent imaging.

$$\sum_{\lambda}\left(\mathcal{F}^{-1}\left\{\mathcal{F}\{\boxed{\cdot}\}\xrightarrow{\text{true object}}\boxed{\triangle}\xrightarrow{\text{OTF}}\boxed{\begin{array}{l}\text{photon \% at}\\ \text{a given }\lambda\end{array}}\right\}\right)\to\boxed{\begin{array}{c}\text{short exposure}\\ \text{image}\end{array}}$$

**Figure 2 Creating a short exposure image from the true object and system model**
Sum the short exposure images of the narrow spectral regions to create a complete short exposure image for the entire bandpass.

3. Results

field data: We reduced the field data sets with a time duration of 12ms. Resulting photon counts per frame were 67, 136, 201, 166, 173, 199, 104, and 124 in order of increasing spectral bandwidth. The single frame SNR curves are shown in figure 3. These are the SNRs of the data with 3000 frames processed; normalization has not been made for the particular number of photons per frame. Notice that the general shape agrees with that expected in that the low spatial frequency seeing spike is present with a subsequent plateau extending toward the diffraction limit. Note also that the midrange values hover around 0.1, but do not appreciably exceed it. Therefore, tOSC's assumption of dominant photon noise applies.

One significant characteristic of these curves is the considerable attenuation of the SNR in the mid and high spatial frequency range for the larger bandwidths. While this appears to directly contradict the increasing SNR theory, in fact, it does not. Fried and Hench's theory relies on the premise that the photon count will rise with the bandwidth; whereas, the photon rate across all of the bandpasses of the field data is relatively constant. Therefore, the theory is not directly applicable to the field data as mentioned earlier. The attenuation at high frequencies is consistent with the expectation that spectral decorrelation becomes significant as the bandwidth rises. The simulation will be used as an intermediary for theory validation.

To ensure the benefits of short exposure, we reduced the data with frames of 6ms, 12ms and 25ms duration to examine the temporal effects. Figure 4 shows the results for the data of 40nm bandwidth. The temporal effects seen here are typical of the other data sets. These results indicate that the SNR drops slightly (10-90%) in the mid and high spatial frequencies. Since the 12ms data is not appreciably different from the 6ms data, we concluded that 12ms was adequate for our purposes since it provides more photons per frame than its 6ms counterpart.
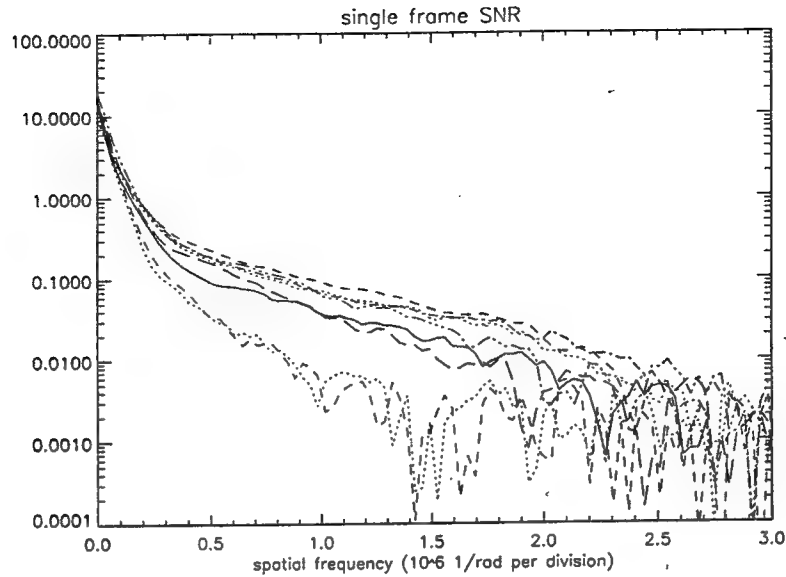
single frame SNR

Figure 3 Single frame SNR of the field data
The single frame power spectrum SNR of the field data over 3000
frames with 12 ms duration. Bandwidth: (—) 30nm,
(. . . [upper]) 40nm, (- - - [upper]) 50nm, (- . -) 70nm
(- ... -) 100nm, (— — —) 140nm, (. . . [lower]) 190nm,
(- - - [lower]) 300nm.
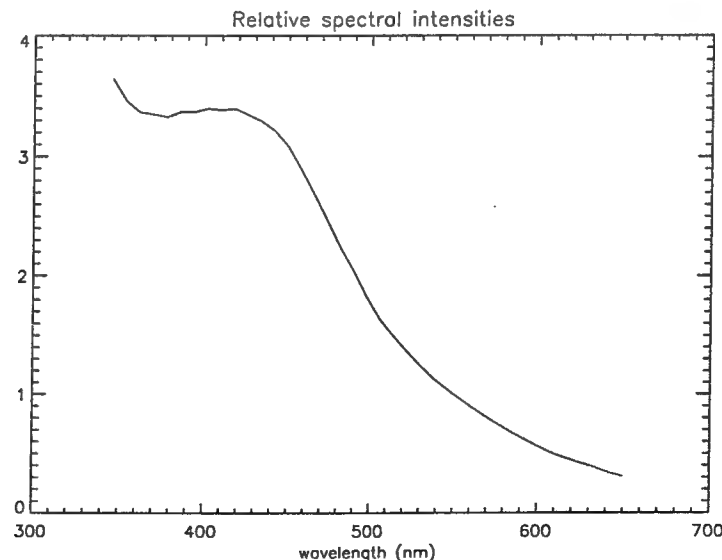

normalized power spectrum
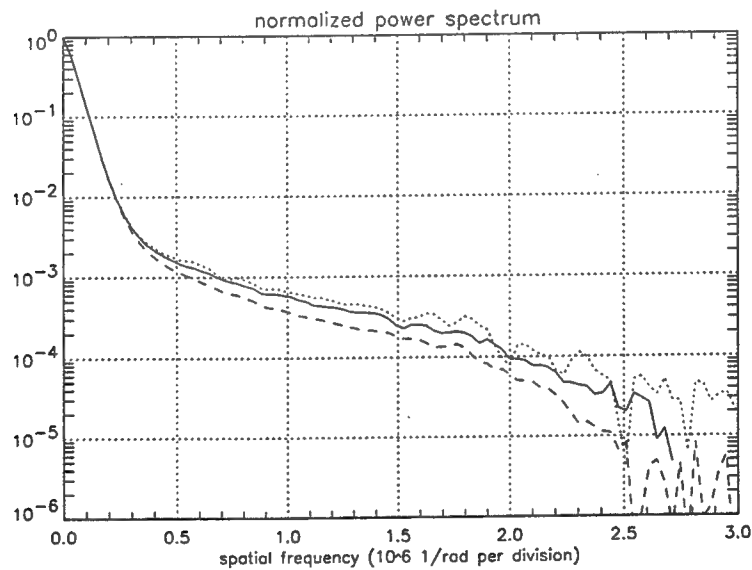
**Figure 4 Temporal comparison of the normalized power
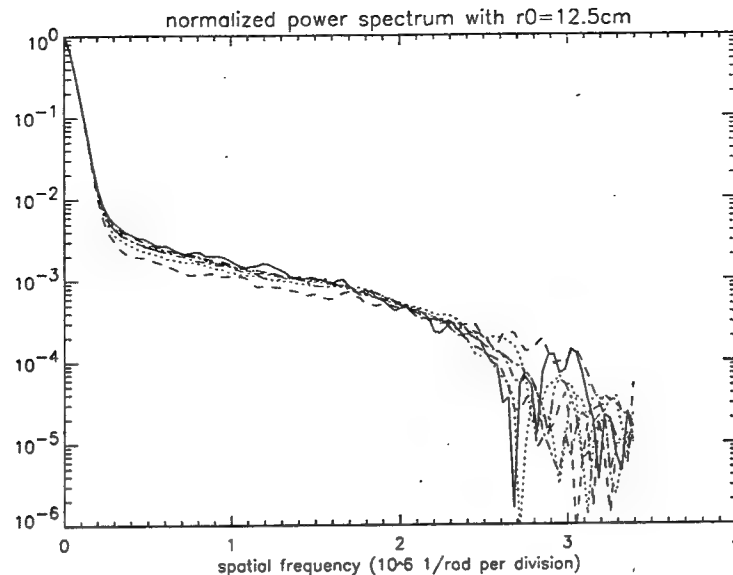spectrum mean**
Radial average of the power spectrum mean of the 40nm
bandpass field data over 3000 frames and normalized to
1 photon/frame. Time duration: (. . .) 6ms, (—) 12ms,
(- - -) 25ms

simulated data: Our first goal of simulation is to accurately model the field data. For this purpose, the simulation allows input of the size of the primary and secondary mirrors, the spectral dependencies, the number of photons per frame and the value of r_0 (the atmospheric effective coherence diameter). The spectral dependencies consist of the star itself, the MAMA detector and the spectral filters. Figure 5 shows the product of the detector spectral efficiency and the black body curve for the Alpha Leo star (i.e., Planck's law, T=13000K). Note that the combined product reveals an intensity at the lowest wavelengths (350nm and 460nm) that is approximately a factor of 10 higher than at the highest wavelength (650nm). This fact must be considered in comparing our results with the increasing SNR theory since Fried and Hench assume that spectral brightness does not vary significantly when making their claim. (1)

To match the field data, each simulated data set that corresponds to a specific spectral filter uses the same photon count as seen in the field data. The value of r_0 was chosen to match the seeing spike of the power spectrum of the simulated data to that of the field data. A value of 12.5cm provides the desired agreement. The power spectrum plots of the simulated data (figure 6) calculated over 1000 frames show some spectral dependence, but not to the extent seen in the field data. For this figure, r_0=12.5cm for all of the bandpasses, although, for the bandpasses above 100nm, the seeing apparently changed because 11cm provides a better match.

Figure 5 Relative spectral intensities of the object and detector.
The product of the Phillips Lab's MAMA detector's spectral efficiency
and the object's spectral irradiance (modeled by Planck's law, T=13000K).

37-10

Figure 6 Normalized power spectrum mean for the simulated data
Radial average of the power spectrum mean over 1000 frames, normalized to 1 photon/frame. Bandwidth: (———) 30nm, (. . . [upper]) 40nm, (- - - [upper]) 50nm, (- . -) 70nm (- ... -) 100nm, (. . . [lower]) 140nm, (- - - [lower]) 190nm

Figure 7 shows the power spectra for both the field data and the simulated data for the 30nm bandpass set. Notice that although the seeing spikes match, the mid and high spatial frequencies of the simulated data are notably higher than the field data. This result, present for each of the bandwidths considered, indicates that there are unmodeled factors present in the field data. A small drop is expected due to the temporal effects mentioned earlier, but that does not account for the majority of the discrepancy. Frequently, this type of attenuation arises due to first order aberrations such as defocus of the telescope, spherical aberrations in the optical path or coma. In inserting these effects into our model, however, we were not able to duplicate the severity of the decline of the magnitude of the power spectrum seen in the field data.

4. Conclusions

While the field data does demonstrate spectral dependence, we are not yet ready to use the data to verify tOSC's theory that increasing spectral bandwidth will never cause a decrease in the power spectrum SNR. (1) First we will need to further analyze the optical path used in the field
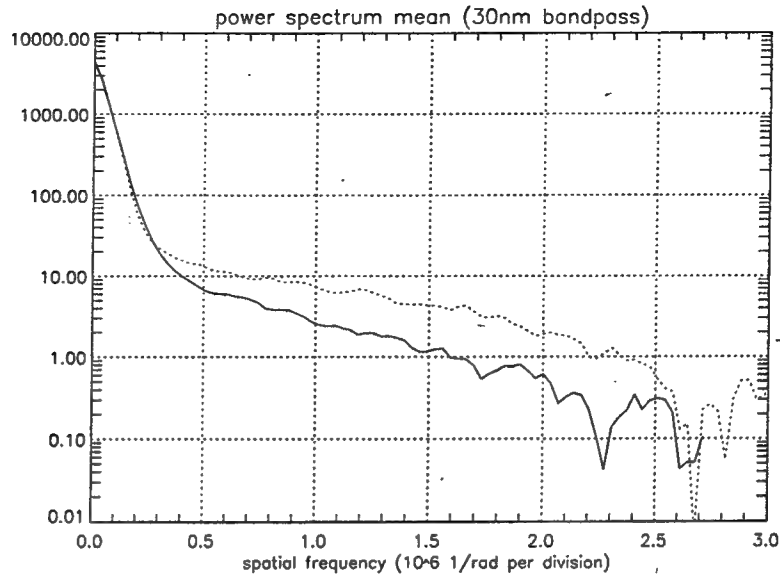
Figure 7 Power spectrum mean of simulated and field data
Radial average of the power spectrum mean of the simulated
data (. . ., 1000 frames) and the field data (——, 3000 frames) for
the 30nm bandpass set.

data collection to more accurately model any aberrations. In addition, we will analyze the data

collection process more closely to ensure that the data collection system did not significantly

alter the photon statistics of the photoevents. When we are able to model the field data more

closely, we will then be able to normalize the field data and subsequently test tOSC's theory.

This theory is limited to the mid range of spatial frequencies. With an accurate model in hand,

we plan to investigate other spatial frequency regions.

REFERENCES

1. D. L. Fried and D.L. Hench, Spectral bandwidth and adaptive optics considerations for white light speckle imagery. *the Optical Sci. Co.*, rep. TR-1033, 1990.
2. A. Labeyrie, Attainment of diffraction-limited resolution in large telescopes by Fourier analyzing speckle patterns in star images. *Astr. Astrophys.* **6,** 85-87 (1970).
3. M. Shoemaker, Frequency Domain Analysis of Short Exposure, Photon-Limited Astronomical Data. AFOSR SFRP Final Report, 1994.

OPTIMAL WARGAMING SCENARIO GENERATION
FOR REAL TIME AUTOMATIC CONTROL
OF MANNED AND UNMANNED
AERIEL VEHICLES

Boris Stilman
Professor
Department of Computer Science and Engineering

University of Colorado at Denver
Campus Box 109
Denver, CO 80217-3364

Final Report for:
Summer Faculty Research Program
Phillips Laboratory

August 1995

OPTIMAL WARGAMING SCENARIO GENERATION FOR REAL TIME AUTOMATIC CONTROL OF MANNED AND UNMANNED AERIEL VEHICLES

Boris Stilman
Professor
Department of Computer Science & Engineering
University of Colorado at Denver

Abstract

We consider the application of Linguistic Geometry tools to the problems of control of remotely piloted aircrafts in conjunction with generation of the optimal combat scenario. These problems are intractable, employing conventional approaches because of the number of variations to be generated and evaluated.

In this report we introduce two examples of such problems. In the first, simplified example, the aircrafts move in a serial mode, one aircraft at a time, and motions of opposing sides alternate. A dramatic search reduction, from a million variations down to 50, achieved in this problem allows us to introduce concurrency inherent to the real world combat. In the second example, all the aircrafts, cooperating and opposing, can move concurrently. This results in a significant growth of the branching factor (up to 300) because all the combinations of simultaneous motions are legal. Another difficulty (in the second example) is that each side when moving is uncertain about the concurrent motions of the opposing side. The Linguistic Geometry tools solved this problem, demonstrating a dramatic reduction of the branching factor, from 324 to 1.52.

Results of the summer research are reflected in 7 research papers. Two of these papers were submitted to research journals and four – to international conferences. Conference papers have been accepted for presentation and publication in USA, France, Turkey, and Greece. One of this papers is co-authored by Capt. Dr. Doug Dyer from the Air Force Phillips Laboratory. This is an invited paper to be presented at the 5th International Conference on Human-Machine Interaction and Artificial Intelligence in Aerospace to be held in Toulouse, France, in September 1995.

Results of the research allowed us to initiate the development of a software prototype of the system for simulation and control of the real world aerospace combat with participation of aircrafts, satellites, and UAVs. This work is currently under way at the Air Force Phillips Laboratory, Kirtland AFB, NM, and will continue in collaboration with the University of Colorado at Denver and Sandia National Laboratories.

OPTIMAL WARGAMING SCENARIO GENERATION FOR REAL TIME AUTOMATIC CONTROL OF MANNED AND UNMANNED AERIEL VEHICLES

Boris Stilman

1. Introduction

A military objective in strategic warfare is to effectively detect missile launches. Although satellites can detect the physical features of a launch, the false-positive error rate is higher than desirable. An effective strategy is to use a reconnaissance aircraft to further investigate launches reported by a satellite's sensors. An advantage of this approach is that an aircraft has the flexibility to respond appropriately, rather than just detect. For example, if a launch site is found, the aircraft can destroy it; if no launch site is found where it was expected, then the aircraft can search in a distance-bounded area for a mobile launcher in transit.

For reconnaissance, aircraft piloted by autonomous agents have certain advantages including less risk to and less utilization of human pilots and non-susceptibility to jamming in the case of remotely piloted aircraft. However, if we are to use an autonomous reconnaissance aircraft, the algorithms which implement it must include those for decision-making while planning a route through territory guarded by aggressive counterparts. The problems described in this paper pit two adversarial aircraft against two autonomous aircrafts, the so-called unmanned aerial vehicles (UAV), on a mission to reconnoiter supposed launch sites. Similar problems of generating and real time replanning of the combat scenarios are essential for the Navy and Army combat.

Problems of this type are usually described mathematically in the form of pursuit-evasion differential games. The classic approach based on the theory of Differential Games is insufficient, especially in case of dynamic, multiagent models (Garcia-Ortiz et al., 1993). It is well known that there exists a small number of differential games, for which exact solutions are available. There are a few more for which numerical solutions can be computed, under rather restrictive conditions, in a reasonable amount of time. It is even worse that each of these games is one-to-one which is very far from the real world combat scenario. They are also of the "zero-sum type" which does not allow a new agent to join the game or all the agents of both sides to be disengaged. Other difficulties arise from the requirements of the 3D modeling and from limitation of the lifetime of the agents.

Following (Rodin, 1988; Shinar, 1990) discrete-event modeling of complex control systems can be implemented as a purely interrogative simulation. These techniques can be based on generating geometrically meaningful states rather than time increments with due respect to the timeliness of actions. By discretizing time, a finite game tree can be obtained. The nodes of the tree represent the states of the game, where the players can select their controls for a given period of time. It is also possible to distinguish the respective moves of the two sides (including simultaneous actions). Thus, the branches of the tree are the moves in the game space. The pruning of such tree is the basic task of heuristic search techniques. Interrogative approach to control problems offers much faster execution and clearer simulator definition (Lirov et al., 1988). Several alternative search methods are applicable to this type of problem. However, in all cases the branching factor for search is unacceptably large unless very good domain knowledge is used to direct search. All known, successful AI programs that address large problems make heavy use of domain information to guide search. Unfortunately, it's difficult to formalize a process for getting the appropriate domain knowledge into the search algorithm so it can do some good. The successes to date have resulted from ad hoc improvement of programs that didn't work as well initially.

The field of efficient control of the multiagent discrete differential games needs new technology

from the science of artificial intelligence.

There are many such problems where human expert skills in reasoning about complex goal-oriented systems are incomparably higher than the level of modern computing systems. Unfortunately, problems of tactics planning and automatic control of autonomous agents such as aerospace vehicles, space stations and robots with cooperative and opposing interests are of the type where human problem-solving skills can not be directly applied. Moreover, there are no highly-skilled human experts in these fields ready to substitute for robots (on a virtual model) or transfer their knowledge to them. There is no grand-master in robot control, although, of course, the knowledge of existing experts in this field should not be neglected – it is even more valuable.

In this respect it is very important to study expert reasoning about similar complex systems in the areas where the results are successful, in order to discover the keys to success, and then apply and adopt these keys to the new, as yet, unsolved problems, and first and foremost to the aerospace critical complex systems.

2 Background
In the beginning of 80's Botvinnik, Stilman, and others developed one of the most interesting and powerful heuristic hierarchical models. It was successfully applied to scheduling, planning, control, and computer chess. The hierarchical networks were introduced in (Botvinnik, 1984; Stilman, 1977) in the form of ideas, plausible discussions, and program implementations. We consider this model as an ideal case for transferring the developed search heuristics to other domains employing formal mathematical tools.

An application of the developed model to a chess domain was implemented in full as program PIONEER (Botvinnik, 1984). Similar heuristic model was implemented for power equipment maintenance in a number of computer programs being used for maintenance scheduling all over the former USSR (Botvinnik et al., 1983; Stilman, 1985, 1993a).

In the 1960's, a formal syntactic approach to the investigation of properties of natural language resulted in the fast development of a theory of formal languages by Chomsky (1963), Ginsburg (1966), and others. This development provided an interesting opportunity for dissemination of this approach to different areas. In particular, there came an idea of analogous linguistic representation of images. This idea was successfully developed into syntactic methods of pattern recognition by Fu (1982), Narasimhan (1966), and Pavlidis (1977), and picture description languages by Shaw (1969), Feder (1971), and Rosenfeld (1979).

Searching for adequate mathematical tools formalizing human heuristics of dynamic hierarchies, we have transformed the idea of linguistic representation of complex real-world and artificial images into the idea of similar representation of complex hierarchical systems (Stilman, 1985). However, the appropriate languages possess more sophisticated attributes than languages usually used for pattern description. The origin of such languages can be traced back to the research on programmed attribute grammars by Knuth (1968), Rozenkrantz (1969).

A mathematical environment (a "glue") for the formal implementation of this approach was developed following the theories of formal problem solving and planning by Nilsson (1980), Fikes and Nilsson (1971), Sacerdoti (1975), McCarthy (1980), McCarthy and Hayes (1969), and others based on first order predicate calculus.

We have applied formal linguistic tools to representation of complex hierarchical systems (Stilman, 1985). To discover the inner properties of human expert heuristics, which have been successful in a certain class of complex control systems, we develop a formal theory, the so-called *Linguistic Geometry* (Stilman, 1993-95). The hierarchy of subsystems is represented as a *hierarchy of formal attribute languages* where each "sentence" (a group of "words" or symbols) of the lower level language corresponds to the "word" of the higher level one.
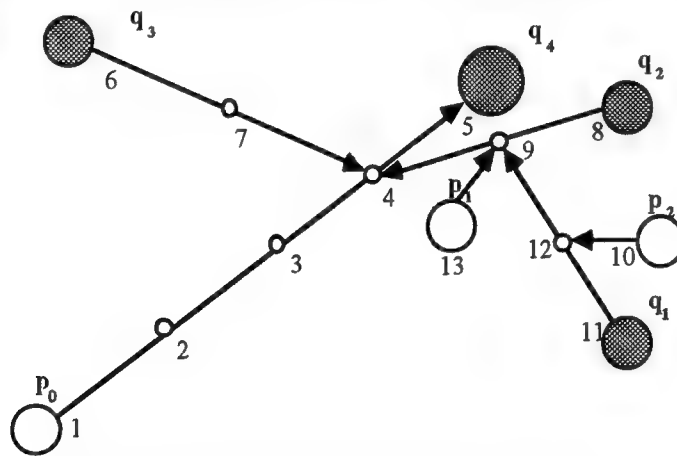
3 Brief Introduction to Linguistic Geometry
The robotic model can be represented as
$$< X, P, R_p, \{ON\}, v, S_i, S_t, TR >,$$
the so-called Complex System (Stilman, 1993a, 1993b). The set X represents the operational

district, which could be the area of combat operation, broken into smaller square areas, "points", e.g., in the form of the big square or grid. It could be a space operation, where X represents the set of different orbits, or an air force battlefield, etc. P is the set of elements (robots or autonomous vehicles). It is broken into two subsets P_1 and P_2 with opposing interests; $R_p(x,y)$, the set relations of reachability, represent moving capabilities of different elements for different problem domains: robot p can move from point x to point y if $R_p(x, y)$ holds. Some of the robots can crawl, others can jump or ride, sail and fly, or even move from one orbit to another. Some of them move fast and can reach point y (from x) in "one step", i.e., $R_p(x, y)$ holds, others can do that in k steps only, and many of them can not reach this point at all. $ON(p)=x$, if robot p is at the point x; $v(p)$ is the value of robot p. This value might be determined by the technical parameters of the robot. It might include the immediate value of this robot for the given combat operation; S_i is an arbitrary initial state of operation for analysis, or the starting state; S_t is the set of target states. These might be the states where robots of each side reached specified points. On the other hand, S_t can specify states where opposing robots of the highest value are destroyed. The set of WFF $\{ON(p_j) = x_k\}$ corresponds to the list of robots with their coordinates in each state. TRANSITION(p, x, y) from TR represents the move of the robot p from the location x to location y; if a robot of the opposing side stands on y, a removal occurs, i.e., robot on y is destroyed and removed.

To find the optimal (suboptimal) operation of the Complex System (according to certain criteria of optimality) we introduce a new representation of state, the Hierarchy of Formal Languages.

Fig. 1. Network language interpretation.

The lowest level of this Hierarchy, Language of Trajectories, is a formal description of the set of various paths between certain locations of the Complex System. An element might follow a path to achieve the goal "connected with the end point of this path."

The next level, Language of Zones, is a formalization of a set of networks of certain paths unified by the mutual goal. For example, in the robot control model a network of planning paths represents a local fight or a draft short-range plan for approaching local goal in hazardous environment, i.e., getting over mobile and immobile obstacles (Stilman, 1993b, 1993c, 1993d, 1994a). An example of such network is shown in Fig. 1. The basic idea behind these networks is as follows. Element p_0 should move along the main trajectory $a(1)a(2)a(3)a(4)a(5)$ to reach the ending point 5 and remove the target q_4 (an opposing element). Naturally, the opposing elements should try to disturb those motions by controlling the intermediate points of the main trajectory. They should come closer to these points (to the point 4 in Fig. 1) and remove element p_0 after its arrival (at point 4). For this purpose, elements q_3 or q_2 should move along the trajectories $a(6)a(7)a(4)$ and $a(8)a(9)a(4)$, respectively, and wait (if necessary) on the next to last point (7 or 9) for the arrival of element p_0 at point 4. Similarly, element p_1 of the same side as p_0 might try to

disturb the motion of q_2 by controlling point 9 along the trajectory $a(13)a(9)$. It makes sense for the opposing side to include the trajectory $a(11)a(12)a(9)$ of element q_1 to prevent this control.

A formal representation of the Zone corresponding to the trajectory network in Fig. 1 is as follows.

$$Z = t(p_0, a(1)a(2)a(3)a(4)a(5), 4)\ t(q_3, a(6)a(7)a(4), 3)$$
$$t(q_2, a(8)a(9)a(4), 3)t(p_1, a(13)a(9), 1)t(q_1, a(11)a(12)a(9), 3)\ t(p_2, a(10)a(12), 1)$$

Assume that the goal of the white side is to remove target q_4, while the goal of the black side is to protect it. According to these goals, element p_0 starts the motion to the target, while black starts in its turn to move elements q_2 or q_3 to intercept element p_0. Actually, only those black trajectories are to be included into the Zone where the motion of the element makes sense, i. e., the *length of the trajectory is less than the amount of time (third parameter τ) allocated to it*. For example, the motion along the trajectories $a(6)a(7)a(4)$ and $a(8)a(9)a(4)$ makes sense, because they are of length 2 and time allocated equals 3: each of the elements has 3 time increments to reach point 4 to intercept element p_0 assuming one would go along the main trajectory without move omission and all the intercepting elements will move *simultaneously* (if necessary). According to definition of Zone, the trajectories of white elements (except p_0) could only be of the length 1, e.g., $a(13)a(9)$ or $a(10)a(12)$. As element p_1 can intercept the motion of the element q_2 at the point 9, Black includes into the Zone the trajectory $a(11)a(12)a(9)$ of the element q_1, which has enough time for motion to prevent this interception. The total amount of time allocated to the whole bundle of black trajectories connected (directly or indirectly) with the given point of the main trajectory is determined by the number of that point. For example, for the point 4, it equals 3 time increments.

The top level of the Hierarchy of Languages, the Language of Translations, is a representation of the search trees generated by the Linguistic Geometry tools.

4   First Robotic Model with Alternating Serial Motions

Robots with different moving capabilities are shown in Fig. 2 . The operational district X is the table 8 x 8. Robot W-FIGHTER (White Fighter) located at h8, can move to any next square (shown by arrows). The other robot B-UAV (Black UAV) from h5 can move only straight ahead, one square at a time, e.g., from h5 to h4, from h4 to h3, etc. Robot B-FIGHTER (Black Fighter) at a6, can move to any next square similarly to robot W-FIGHTER (shown by arrows). Robot W-UAV located at c6 is analogous with the robot B-UAV; it can move only straight ahead but in reverse direction.
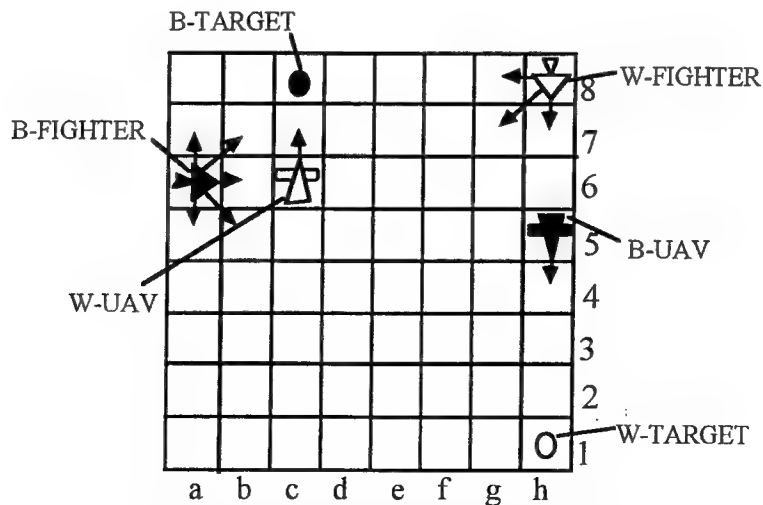

Fig. 2. A problem for robotic vehicles.

Thus, robot W-FIGHTER on h8 can reach any of the points $y \in \{h7, g7, g8\}$ in one step, i.e., $R_{W\text{-FIGHTER}}(h8, y)$ holds, while W-UAV can reach only c8 in one step.

Assume that robots W-FIGHTER and W-UAV belong to one side, while B-FIGHTER and B-UAV belong to the opposing side: W-FIGHTER $\in P_1$, W-UAV $\in P_1$, B-FIGHTER $\in P_2$, B-UAV $\in P_2$. Also assume that two more robots, W-TARGET and B-TARGET, (launch sites or mobile launchers) are located at h1 and c8, respectively. In this problem we consider possible motions of launchers as bounded within the areas h1 and c8, respectively.

W-TARGET belongs to P_1, while B-TARGET $\in P_2$. Each of the UAVs can destroy the TARGET ahead of the course; it also has powerful weapons capable to destroy opposing FIGHTERs on the next diagonal squares ahead of the course. For example W-UAV from c6 can destroy opposing FIGHTERs on b7 and d7. Each of the FIGHTERs is capable to destroy an opposing UAV approaching its location, but it also capable to protect its friendly UAV approaching its prospective location. In the latter case the joint protective power of the combined weapons of the friendly UAV and FIGHTER can protect the UAV from interception. For example, W-FIGHTER located at d6 can protect W-UAV on c6 and c7.

The battlefield considered can be broken into two local operations. The first operation is as follows: robot B-UAV should reach point h1 to detect and destroy the W-TARGET, while W-FIGHTER will try to intercept this motion. The second operation is similar: robot W-UAV should reach point c8 to destroy the B-TARGET, while B-FIGHTER will try to intercept this motion. After destroying the opposing TARGET the attacking side is considered as a winner of the local operation and the global combat. The only chance for the opposing side to avenge itself is to hit its TARGET on the next time increment and this way end the battle in a draw. The conditions considered above give us S_t, the description of target states of the Complex System. The description of the initial state S_i is obvious and follows from Fig. 2 .

Assume that due to the shortage of resources (which is typical in real combat operation) or some other reasons, each side can not participate in both operations simultaneously. It means that during the current time increment, in case of White turn, either W-UAV or W-FIGHTER can move. Analogous condition holds for Blacks. Of course, it does not mean that if one side began participating in one of the operations it must complete it. Any time on its turn each side can switch from one operation to another, e.g., transferring resources, and later switch back. These constraints convert our problem into the air combat problem with *serial* motions. In Section 6 we will consider different model with complete relaxation of these requirements.

It seems that local operations are independent, because they are located far from each other. Moreover, the operation of B-UAV from h5 looks like unconditionally winning operation, and, consequently, the global combat can be easily won by the Black side. *Is there a strategy for the White side to make a draw?*

Of course, this question can be answered by the direct search employing, for example, minimax algorithm with alpha-beta cut-offs. Experiments with the computer programs showed that in order to solve this problem employing conventional approaches the search tree should include about a million moves (transitions).

5 Scenario Generation for Robotic Model: Alternating Serial Motions

Consider how the hierarchy of languages works for the optimal control of the Robotic System introduced above (Fig. 2).

We generate a string of the Language of Translations representing it as a conventional search tree (Fig. 3) and comment on its generation.

First, the Language of Zones in the start state is generated. The targets for attack are determined within the limit of five steps. It means that horizon H of the language $L_Z(S)$ is equal to 5, i.e., the length of main trajectories of all Zones must not exceed 5 steps. All the Zones generated in the start state are shown in Fig. 4. Zones for FIGHTERs as attacking elements are shown in the left diagram, while Zones for UAVs – in the right one. For example, one of the Zones for W-UAV,

Z_{WU} is as follows:
$$\text{Z}_{\text{WU}} = t(P, \boldsymbol{a}(c6)\boldsymbol{a}(c7)\boldsymbol{a}(c8), 2)t(K, \boldsymbol{a}(a6)\boldsymbol{a}(b7)\boldsymbol{a}(c8), 3)$$
$$t(K, \boldsymbol{a}(a6)\boldsymbol{a}(b7)\boldsymbol{a}(c7), 2)t(P, \boldsymbol{a}(c6)\boldsymbol{a}(b7), 1)$$

The second trajectory of B-FIGHTER $\boldsymbol{a}(a6)\boldsymbol{a}(b6)\boldsymbol{a}(c7)$ leading to the square c7 is included into different Zone; for each Zone only one trajectory from each bundle of trajectories is taken.

Generation begins with the move 1. c6-c7 in the White Zone with the target of the highest value and with the shortest main trajectory. The order of consideration of Zones and particular trajectories is determined by the grammar of translations. The computation of move-ordering constraints is the most sophisticated procedure in this grammar. It takes into account different parameters of Zones, trajectories, and the so-called chains of trajectories.
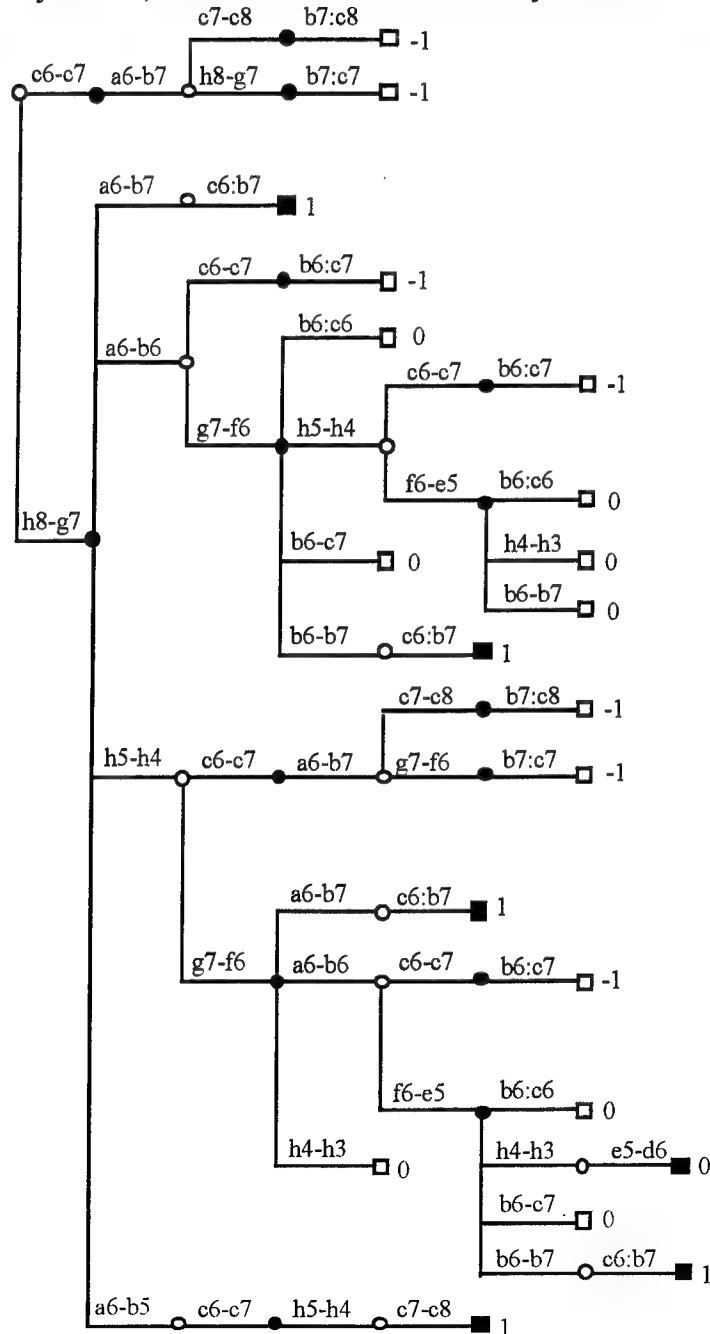
Fig. 3. Search tree for the robotic vehicles with alternating serial motions.
Next move, 1. ... a6-b7, is in the same Zone along the first negation trajectory. Interception

continues: 2. c7-c8 b7:c8 (Fig. 5, left).

Symbol "." means the removal of element. Here the grammar cuts this branch with the value of -1 (as a win of the Black side). This value is given by the special procedure of "generalized square rules" built into the grammar.

Then, the grammar initiates the backtracking climb. Each backtracking move is followed by the inspection procedure, the analysis of the subtree generated in the process of the earlier search. After climb up to the move 1. ... a6-b7, the tree to be analyzed consists of one branch (of two plies): 2. c7-c8 b7:c8. The inspection procedure determined that the current minimax value (-1) can be "improved" by the improvement of the exchange on c8 (in favor of the White side). This can be achieved by participation of W-FIGHTER from h8, i.e., by generation and inclusion of the new so-called "control" Zone with the main trajectory from h8 to c8. The set of different Zones from h8 to c8 (the bundle of Zones) is shown in Fig. 5, right. The move-ordering procedure picks the subset of Zones with main trajectories passing g7. These trajectories partly coincide with the main trajectory of another Zone attacking the opposing W-UAV on h5. The motion along such trajectories allows to "gain time", i.e., to approach two goals simultaneously.
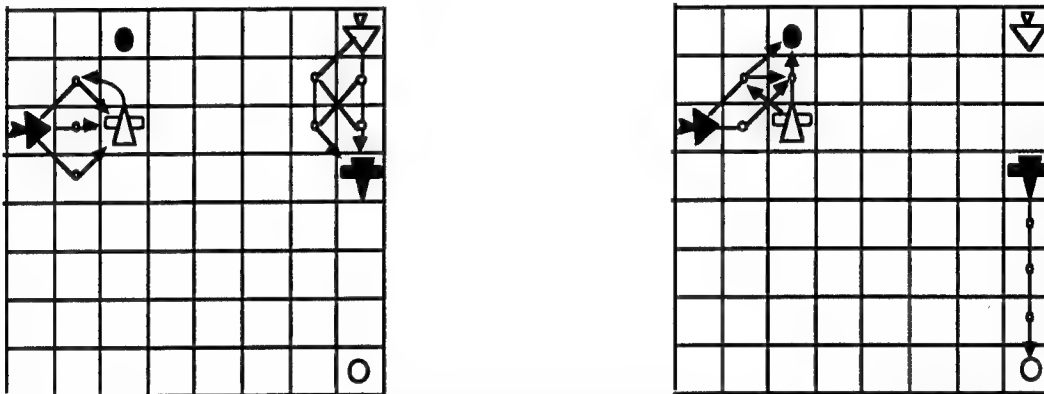

Fig. 4. Interpretation of the Zones in the initial state of the Robot Control Model.
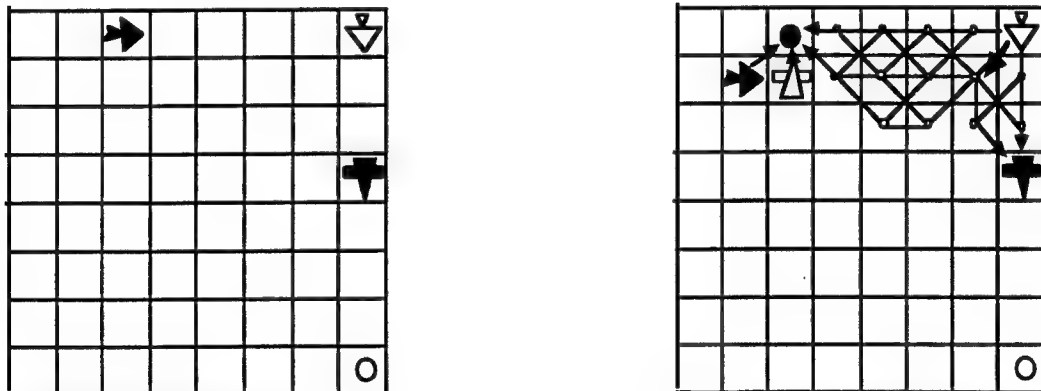

Fig. 5. States where the control Zone from h8 to c8 was detected (left) and where it was included into the search (right)

The generation continues: 2. h8-g7 b7:c7. Again, the procedure of "square rules" cuts the branch, evaluates it as a win of the black side, and the grammar initiates the climb. Analogously to the previous case, the inspection procedure determined that the current minimax value (-1) can be improved by the improvement of the exchange on c7. Again, this can be achieved by the inclusion of Zone from h8 to c7. Of course, the best "time-gaining" move in this Zone is 2. h8-g7, but it was already included (as move in the Zone from h8 to c8), and it appeared to be useless. No other branching at this state is generated.
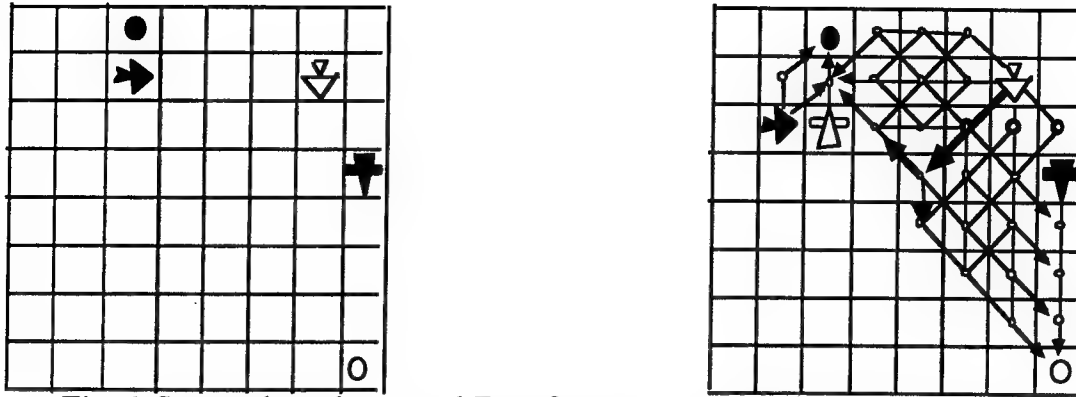
Fig. 6. States where the control Zone from g7 to c7 was detected (left) and where it was included into the search (right).

The inspection procedure does not find new Zones to improve the current minimax value, and the climb continues up to the start state. The analysis of the subtree shows that inclusion of Zone from h8 to c8 in the start state can be useful: the minimax value can be improved. Similarly, the most promising "time-gaining" move is 1. h8-g7. The Black side responded 1. ... a6-b7 along the first negation trajectories a(a6)a(b6)a(c7) and a(a6)a(b6)a(c8) (Fig. 4, right). Obviously, 2. c6:b7, and the branch is terminated. The grammar initiates the climb and move 1. ... a6-b7 is changed for 1. ... a6-b6 along the trajectory a(a6)a(b6)a(c7). Note, that grammar "knows" that in this state trajectory a(a6)a(b6)a(c7) is active, i.e., B-FIGHTER has enough time for interception. The following moves are in the same Zone of W-UAV: 2. c6-c7 b6:c7. This state is shown in Fig. 6, left. The "square rule procedure" cuts this branch and evaluates it as a win of the Black side.

New climb up to the move 2. ... a6-b6 and execution of the inspection procedure resulted in the inclusion of the new control Zone from g7 to c7 in order to improve the exchange on c7. The set of Zones with different main trajectories from g7 to c7 is shown in Fig. 6, right. Besides that, the trajectories from g7 to h4, h3, h2, and h1 are shown in the same Fig. 6. These are "potential" first negation trajectories. It means that beginning with the second symbol a(f6),a(g6) or a(h6) these trajectories become *first negation* trajectories (Section 6) in the Zone of B-UAV h5. Speaking informally, from squares f6, g6, and h6 W-FIGHTER can intercept B-UAV (in case of white move). The move-ordering procedure picks the subset of Zones with the main trajectories passing f6. These trajectories partly coincide with the potential first negation trajectories. The motion along such trajectories allows to gain time, i.e., to approach two goals simultaneously. Thus, 2. g7-f6.

This way proceeding with the search we will generate the tree that consists of 48 moves.

6. Second Robotic Model: Total Concurrency

A new robotic model is shown in Fig. 8 (compare with Section 4). The operational district X is the table 8 x 8. Robot W-FIGHTER (White Fighter) located at h8, can move to any next square (shown by arrows). The other robot B-UAV (Black UAV) from h7 can move only straight ahead, one square at a time, e.g., from h7 to h6, from h6 to h5, etc. Robot B-FIGHTER (Black Fighter) at a6, can move to any next square similarly to robot W-FIGHTER (shown by arrows). Robot W-UAV (White UAV) located at c6 is analogous with the robot B-UAV; it can move only straight ahead but in reverse direction. Thus, robot W-FIGHTER on h8 can reach any of the points y $\in$ {h7, g7, g8} in one step, i.e., $R_{\text{W-FIGHTER}}$(h8, y) holds, while W-UAV can reach only c7 in one step.

As in Section 4, robots W-FIGHTER and W-UAV belong to one side, while B-FIGHTER and B-UAV belong to the opposing side: W-FIGHTER $\in P_1$, W-UAV $\in P_1$, B-FIGHTER $\in P_2$, B-UAV $\in P_2$. Also, assume that two more robots, W-TARGET and B-TARGET, (launch sites or mobile launchers) located at h2 and c8, respectively. They can not exit these squares. W-TARGET belongs to P_1, while B-TARGET $\in P_2$. Each of the UAVs can destroy the TARGET ahead of the

38-10

course. Each of the FIGHTERs is able to destroy an opposing UAV approaching its location, but it is also able to destroy an opposing UAV if this UAV itself arrives at the current FIGHTER's location. For example, if the B-FIGHTER is at location c8 and W-UAV arrives there (unprotected) then during the same time increment it destroys the TARGET and is destroyed itself by B-FIGHTER. Each UAV can be protected by its friendly FIGHTER if the latter approached the UAV's prospective location. In this case the joint protective power of the combined weapons of the friendly UAV and FIGHTER can protect the UAV from interception. For example, W-FIGHTER located at d6 can protect W-UAV on c6 and c7.
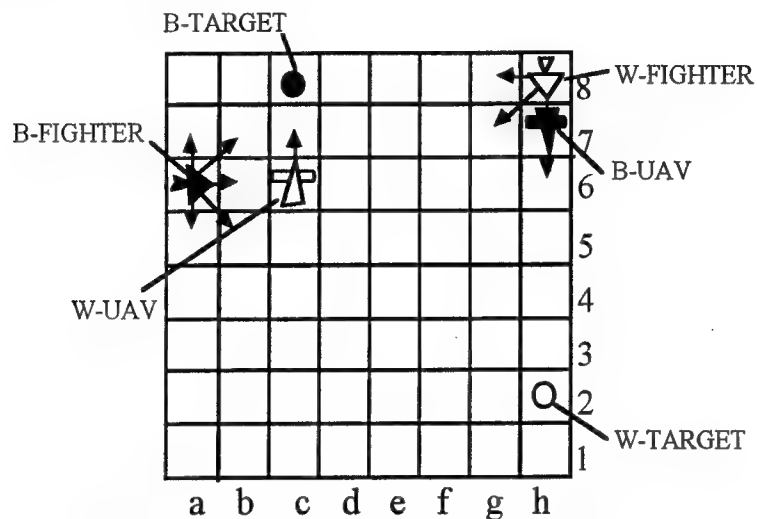
Fig. 8. 2D optimization problem for robotic vehicles with totally concurrent motions.

Each of the UAVs is vulnerable not only to a FIGHTER's attack but also to the explosion of another UAV. If W-FIGHTER hits B-UAV while the latter is fully armed, i.e., it is not at its final destination – square h2, and W-UAV is moving during the same time increment, it will be destroyed as a result of the B-UAV's explosion. If W-UAV is not moving at this moment it is safe. Similar condition holds for B-UAV: it can not move at the moment when W-UAV is being destroyed (excluding c8).

The combat considered can be broken into two local operations. The first operation is as follows: robot B-UAV should reach point h2 to destroy the W-TARGET, while W-FIGHTER will try to intercept this motion. The second operation is similar: robot W-UAV should reach point c8 to destroy the B-TARGET, while B-FIGHTER will try to intercept this motion. After destroying the opposing TARGET and keeping the UAV safe, the attacking side is considered as a winner of the local operation and the global combat. The only chance for the opposing side to avenge is to hit its TARGET at the same time increment and this way end the battle in a *draw*. The conditions considered above give us S_t, the description of target states of the Complex System. The description of the initial state S_i is obvious and follows from Fig. 8.

Assume that all the agents can move *simultaneously*. With respect to our previous example of robotic system (Sections 4, 5) in this model there is no alternation of turns. It means, for example, that during the current time increment, all four vehicles, W-UAV, W-FIGHTER, B-UAV, and B-FIGHTER, three of them, two, one, or none of them can move. Hence, this is a model with *incomplete information* about the current move (before it is done). When moving, each side does not know the opposing side component of the concurrent move, i.e., the immediate opposing side motions, if they are not constrained to one or zero motions and, thus, can be predicted. Moreover, after developing a strategy each side can not follow it because of the uncertainty with the other side current motions. However, if the strategy includes only variations of concurrent moves with single "universal" component (group of motions) for one side good for all possible components of the other side, this strategy can be actually implemented.

It seems that local operations (Fig. 8) are independent, because they are located far from each other. Moreover, the operation of B-UAV from h7 looks like unconditionally winning operation, and, consequently, the global battle can be easily won by the Black side. **Is there a strategy for the White side to make a draw?**

7 Scenario Generation: Total Concurrency and Uncertainty

Consider how the hierarchy of languages works for the optimal control of this model. We have to generate the Language of Trajectories and the Language of Zones in each state of the search. The details of these generations are considered in (Stilman, 1993b, 1993c, 1993d, 1994a, 1994b).

We generate a string of the Language of Translations (Stilman, 1994a) representing it as a search tree (Fig. 9) and comment on its generation. This tree is different from conventional search trees and from the tree in Fig. 3.

Fig. 9. Search tree for the Totally Concurrent Model.

Every concurrent move is represented by *two consecutive arcs*. The arc outgoing the white node represents the White component of a concurrent move, the concurrent motions of the White side,

while the arc outgoing the next black node represents the Black component of the same move.

First, the Language of Zones in the start state is generated. Every agent tries to attack every opposing side agent. The targets for attack are determined within the limit of five steps. It means that horizon H of the language $L_Z(S)$ is equal to 5, i.e., the length of main trajectories of all Zones must not exceed 5 steps. The algorithm for choosing the right value of the horizon is considered in (Stilman, 1994c). All the Zones generated in the start state are shown in Fig. 10. Zones for FIGHTERs as attacking elements are shown in the left diagram, while Zones for UAVs – in the right one.

Generation begins with the concurrent move 1. c6-c7 a6-b7 in the White Zone with the vulnerable Black target of the highest value and the shortest main trajectory. The order of consideration of Zones and particular trajectories is determined by the grammar of translations.

The Black component of this move, 1. ... a6-b7, is in the same Zone along the intercepting trajectory. The interception continues: 2. c7-c8 b7-c8/h7-h6 (Fig. 11, left). This is a triple move. During the second time increment W-UAV hit the TARGET at c8 and was destroyed by the B-FIGHTER at c8. Also, immediately, the attack Zone of the B-UAV from h7 to h2 was activated: h7-h6 is the motion during the same time increment. Here the grammar terminates this branch with the value -1 (as a win of the Black side). This value is given by the special branch termination procedure built into the grammar. This procedure determined that W-FIGHTER is out of the Zone of B-UAV, thus, it can not intercept B-UAV which means that the latter will successfully hit the TARGET on h2.
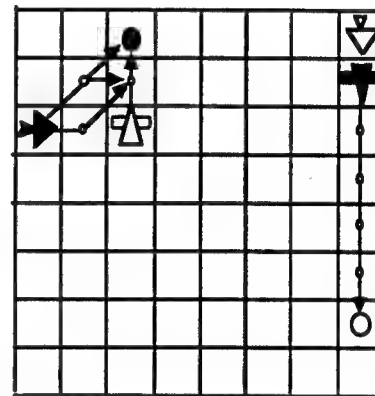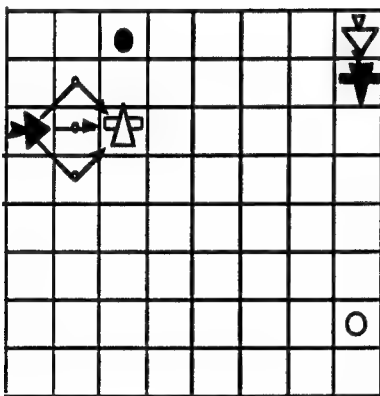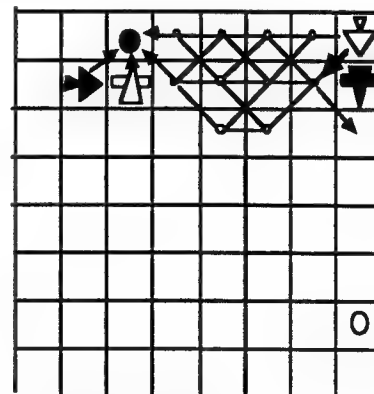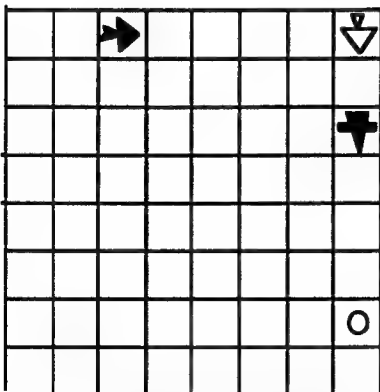

Fig. 10. Zones in the initial state

Fig. 11. States where the control Zones from h8 to c8 were
detected (left) and where they were included into the search (right)

Then, the grammar initiates the backtracking climb. Each backtracking move is followed by the inspection procedure, the analysis of the subtree generated so far (with the current node as a root). After climb up to the move 1. c6-c7 a6-b7, the subtree to be analyzed consists of one branch (of

one move): 2. c7-c8 b7-c8/h7-h6. The inspection procedure determined that the current minimax value (-1) can be "improved" by the improvement of the exchange on c8 (in favor of the White side).
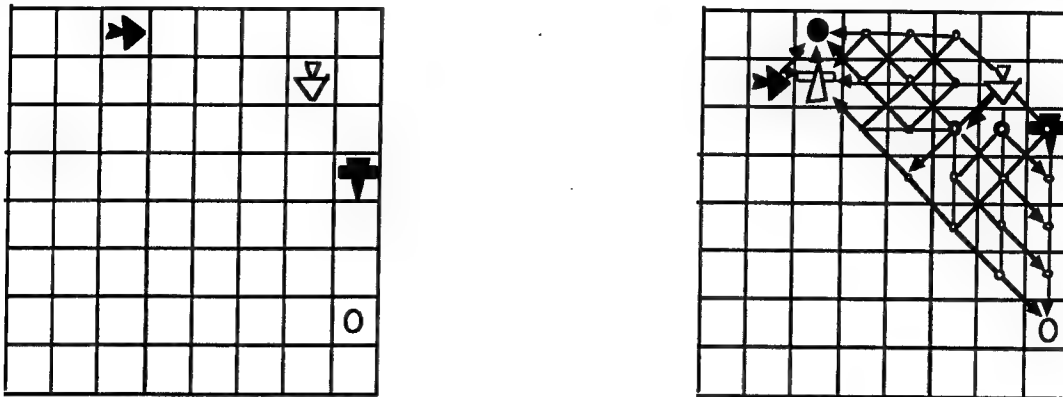

Fig. 12. States where the control Zones from g7 to c7, c8 were detected (left) and where they were included into the search (right)
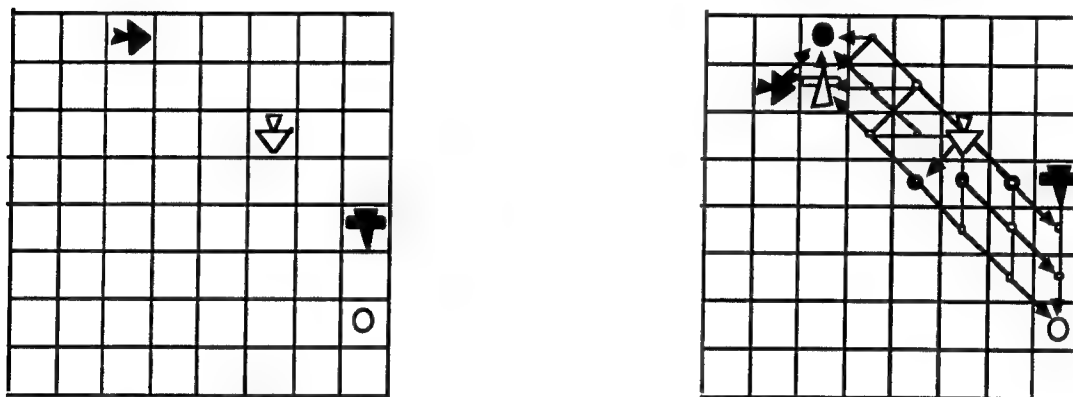

Fig. 13. States where the control Zones from f6 to c7, c8 were detected (left) and where they were included into the search (right).

This can be achieved by participation of W-FIGHTER from h8, i.e., by generation and inclusion of the new so-called "control" Zones with the main trajectory from h8 to c8. These Zones were detected (within the horizon 5) in the terminal state after the move 2. c7-c8 b7-c8/h7-h6, Fig. 11 (left). Obviously they could not be detected in the initial state of this problem (Fig. 10) because the main element, W-UAV, could not "see" the target, B-FIGHTER, within given horizon. Also, at the moment of detection it was too late to include them into the search. These Zones have been stored and kept idle for possible activation at the higher levels of the search tree. The set of different Zones from h8 to c8 (the bundle of Zones) is shown in Fig. 11 (right). The move-ordering procedure picks the subset of Zones with main trajectories passing g7. These trajectories partly coincide with the main trajectory of another Zone attacking the opposing W-UAV on its future location h6. The motion along such trajectories allows to "gain time", i.e., to approach two goals simultaneously.

The generation continues with the simultaneous motion of all four agents, the four-move, W-UAV, W-FIGHTER and B-FIGHTER, B-UAV, in their respective Zones: 2. c7-c8/h8-g7 b7-c8/h7-h6. The B-FIGHTER intercepted W-UAV at c8 while W-FIGHTER is unable to intercept the B-UAV during its attack from h6 to h2. The branch termination procedure determined that W-FIGHTER is outside the B-UAV's attack Zone, terminated this branch, evaluated it as a win for the Black (-1), and initiated the backtracking climb. Move 2. ... was changed for the triple move 2. h8-g7 b7-c8/h7-h6 in attempt to find a better combination of White motions.

Black side, after finding b7-c8/h7-h6 to be a "good" component of the concurrent move 2. in

the previous branches, continues to include this component in the following branches. Obviously, this component is very important. As it was noted above, a totally concurrent model is a model with incomplete information. Each side knows all the previous moves, the history of operation, and, theoretically, all possible future outcomes of the current move, the look-ahead tree. The only thing it does not know is the concurrent action of the opposing side as a component of the current move. Thus, for each side it is important to find not just a good own component of a concurrent move but a component to be good for all components of the opposing side. Such component would allow to avoid uncertainty in constructing an optimal variation, a branch, which can be implemented. A component b7-c8/h7-h6 is a candidate to be a good one for the Black while h8-g7 is a candidate for White.

After 2. h8-g7 b7-c8/h7-h6 termination procedure did not terminate the branch, and the grammar continued 3. c7-c8 h6-h5 in the same Black and White Zones. Then it terminated the branch and evaluated it as a win (-1) for the Black side (Fig. 12, left). Indeed, W-UAV hit B-TARGET on c8 but it is destroyed itself by B-FIGHTER which was waiting for it at c8. Also, W-FIGHTER again is out of the attack Zone of B-UAV from h5 to h2. In this state a set of new control Zones of W-FIGHTER from g7 to c8 were detected and stored as idle to be activated later if necessary.

New climb up to the move 2. h8-g7 b7-c8/h7-h6 and execution of the inspection procedure resulted in the inclusion of the groups of new control Zones from g7 to c7 and c8 in order to improve the exchanges at these locations. Both groups of Zones (to c7 and c8) have been detected earlier in the search tree. The set of Zones with different main trajectories from g7 to c7 and from g7 to c8 is shown in Fig. 12 (right). Besides that, the trajectories from g7 to h4, h3, and h2, are shown in the same Fig. 12. These are "potential" first negation trajectories. It means that beginning with the second symbol a(f6), a(g6) or a(h6) these trajectories become first negation trajectories in the Zone of B-UAV on h6. Speaking informally, from the squares f6, g6, and h6, Zone gateways, W-FIGHTER can intercept B-UAV. The move-ordering procedure picks the subset of Zones with the main trajectories passing f6. These trajectories partly coincide with the potential intercepting trajectories. The motion along such trajectories allows to "gain time", i.e., to approach two goals simultaneously.

Thus, the new White component 3. c7-c8/g7-f6 is included with the same Black component 3. ... h6-h5, the branch was terminated with the value -1. The following climb and branching with inclusion of g7-f6 as a single motion component resulted in 3. g7-f6 h6-h5, and the branch is not terminated. It continues with the move 4. c7-c8 h5-h4. This state is shown in Fig. 13, left. Then this branch is terminated with the value -1. As usual, this value was assigned by the termination procedure which detected that W-FIGHTER is outside the Zone of B-UAV and, thus, does not have enough time for interception.

After the climb, the grammar continued branching 4. c7-c8/f6-e5 h5-h4. The component f6-e5 is selected by the move ordering procedure as the time-gaining move approaching two goals simultaneously, c7 as a goal of the control Zone of W-FIGHTER and one of the gateways (e5, f5, g5) of the Zone of B-UAV (Fig. 13, right). It was also terminated with the value -1. After 4. f5-e5 h4-h3, 5. e5-d6 h4-h3, and 6. c7-c8/d6-d7 h3-h2, the branch is terminated with the value of 0.

It seems that the sought draw is found. The following climb with activation of the inspection procedure at every node ended at the top level. All the attempts of the Black to change the components 4. ... h5-h4, 3. ... h6-h5, 2. ... h7-h6 for a different motion failed. If B-UAV's motion is not included in these concurrent moves the W-FIGHTER appears in the B-UAV's attack Zone and these branches should be terminated with the value 0 which does not improve the current minimax value for Black.

The Black component of 1. c6-c7 a6-b7 was changed for the double motions 1. c6-c7 a6-b7/h7-h6. It seems that this move almost depreciated previous search. The minimax value brought to the top of the subtree outgoing this move is -1. However, the tree generation followed after the change of 1. c6-c7 a6-b7/h7-h6 for the four-move 1. c6-c7/h8-g7 a6-b7/h7-h6 showed that previous search was very important. As a result of this search the grammar *learned key networks*, Zones of W-FIGHTER with main trajectories from h8 to c8, from g8 to c8, and from g7, f6 to c7 and c8. The optimal branch is shown in Fig. 9 with bold lines.

8   Discussion

In case of the model with alternating serial motions (Sections 4, 5) the total number of moves included into the search tree is 48. Obviously, this is a dramatic reduction in comparison with a million-move trees generated by conventional search procedures. The maximum depth of this search is 7, so the branching factor (Nilsson, 1980) is about 1.5. This means that the search is highly goal-oriented.

In case of the totally concurrent model (Sections 6,7) the number of moves included in the search tree is 34. The maximum depth reached is 6. This means that the branching factor of this tree is 1.53. The average number of legal motions for each side, i.e., the average number of different legal components of every concurrent move, is 18. Thus, the average number of legal moves in each state (unreduced branching factor) is $18 \times 18 = 324(!)$, taking into account all the combinations of legal components. As it was in the first example, 34 is a dramatic reduction in comparison with a 324^6 move tree that would have to be generated by conventional search procedures, or even with the theoretical minimum of the minimax search with alpha-beta cut-offs $(324^6)^{1/2} = 18^6 \approx 34$ million.

Search reduction achieved in the serial case with one-at-a-time alternating motions of every vehicle multiplied tremendously in the example with the allowance of *concurrent* moves. The next step of our research will be a formal investigation of the complexity of the hierarchy of languages which represents each state in the search process. It is easy to speculate that the growth from the serial case to the concurrent one is limited by multiplication by a constant factor close to one.

Simplified aerospace combat scenarios considered here are still very close to the original board game problem domain. The power of Linguistic Geometry goes far beyond these limits. The definition of Complex System (Section 3) is generic enough to cover a variety of different complex problem domains. The core component of this definition is the triple X, P, and R_p. Thus, looking at the new problem domain we have to define X, the finite set of points – locations of elements. We do not impose any constraints on this set while the operational district X considered in this paper as well as the original game board have different extra features, e.g., 2D or 3D space connectivity, which is totally unimportant for these problems. Thus, for example, we can consider X as a set of orbits where the elements are in permanent motion with respect to each other. The moving capabilities of mobile units, elements P, in our examples, i.e., the binary relations R_p, are non-sophisticated. For various applications they can be advanced significantly. This is exactly the place for introduction of the variable speed, the gravity impact, the engine impulse duration, etc.

A dramatic search reduction achieved in the serial and concurrent cases and the generic power of Linguistic Geometry tools allowed us to initiate the development of a prototype of the system for simulation and control of the real world aerospace combat with participation of aircrafts, satellites, and UAVs. This work is currently under way at the Air Force Phillips Laboratory, Kirtland AFB, NM, USA, and will continue in collaboration with the University of Colorado at Denver and Sandia National Laboratories.

References

Botvinnik, M.M. (1984). *Computers in Chess: Solving Inexact Search Problems.* Springer Series in Symbolic Computation, New York: Springer-Verlag.

Botvinnik, M., Petriyev, E., Reznitskiy, A., et al. (1983). Application of New Method for Solving Search Problems For Power Equipment Maintenance Scheduling. *Economics and Mathematical Methods* (1030-1041), 6, (in Russian).

Chomsky, N. (1963). Formal Properties of Grammars. in *Handbook of Mathematical Psychology*, eds. R.Luce, R.Bush, E. Galanter., vol. 2 (323-418). New York: Wiley.

Feder, J. (1971). Plex languages. *Inform. Sciences*, 3,(225–241).

Fikes, R.E. and Nilsson, N.J. (1971). STRIPS: A New Approach to the Application of Theorem Proving in Problem Solving. *Artificial Intelligence* 2 (189–208).

Fu, K.S. (1982). *Syntactic Pattern Recognition and Applications*, Prentice Hall, Englewood Cliffs.

Garcia-Ortiz, A. et al. (1993). Application of Semantic Control to a Class of Pursue-Evader Problems, *Int. J. Computers and Mathematics with Appl.*, 26(5), (97-124).

Ginsburg, S. (1966). *The Mathematical Theory of Context-Free Languages*, McGraw Hill, New York.

Knuth, D.E. (1968). Semantics of Context-Free Languages. *Mathematical Systems Theory*, (127–146), 2.

Lirov Y., Rodin, E.Y., McElhaney, B.G., and Wilbur, L.W. (1988). Artificial Intelligence Modeling of Control Systems, *Simulation*, (12-24), 50(1).

McCarthy, J. (1980). Circumscription–A Form of Non-Monotonic Reasoning. *Artificial Intelligence*, (27-39), 13.

McCarthy, J. and Hayes, P.J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence* (463–502), 4.

Narasimhan, R.N. (1966). Syntax–Directed Interpretation of Classes of Pictures. *Comm. of ACM* (166–173), 9.

Nilsson, N.J. (1980). *Principles of Artificial Intelligence*, Palo Alto, CA: Tioga Publ.

Pavlidis, T. (1977). *Structural Pattern Recognition*, New York: Springer-Verlag.

Rodin E. (1988). Semantic Control Theory, *Applied Mathematical Letters*, (73-78), 1(1).

Rodin E., Garcia-Ortiz et al. (1993). Application of Semantic Control to a Class of Pursue-Evader Problems, *Computers and Mathematics with Applications*, 26(5).

Rosenfeld, A. (1979). *Picture Languages, Formal Models for Picture Recognition*, Academic Press.

Rozenkrantz, D.J. (1969). Programmed Grammars and Classes of Formal Languages, *J. of ACM* (107–131), 1.

Sacerdoti, E.D. (1975). The Nonlinear Nature of Plans, *Proc. Int. Joint Conference on Artificial Intelligence.*

Shaw, A.C. (1969). A Formal Picture Description Scheme as a Basis for Picture Processing System, *Information and Control* (9-52), 19.

Shinar, J., (1990). Analysis of Dynamic Conflicts by Techniques of Art. Intelligence, INRIA Report, Antipolis.

Stilman, B. (1977). The Computer Learns. in Levy, D., *1976 US Computer Chess Championship* (83-90). Computer Science Press, Woodland Hills, CA, **USA**.

Stilman, B. (1985). Hierarchy of Formal Grammars for Solving Search Problems. In *Artificial Intelligence. Results and Prospects, Proc. of the Int. Workshop* (63-72), Moscow, **Russia**, (in Russian).

Stilman, B. (1993a). A Linguistic Approach to Geometric Reasoning, **Int. J.** *Computers and Mathematics with Applications* (29-57), 26(7).

Stilman, B. (1993b). Network Languages for Complex Systems, **Int. J.** *Computers and Mathematics with Applications* (51-79), 26(8).

Stilman, B. (1993c). Syntactic Hierarchy for Robotic Systems, *Integrated Comp.-Aid. Engineering,* **(Int. J.)**,(57-81), 1(1).

Stilman, B. (1993d). A Formal Language for Hierarchical Systems Control, *Languages of Design* (**Int. J.**), (333-356), 1(4).

Stilman, B. (1994a). Translations of Network Languages. **Int. J.** *Computers and Mathematics with Applications* (65-98), 27(2).

Stilman, B. (1994b). A Formal Model for Heuristic Search. *Proc. of the 22nd Annual ACM Computer Science Conf.*, (380-389), March 8-10, Phoenix, AZ, **USA**.

Stilman, B. (1994c). Heuristic Networks for Space Exploration, *Telematics and Informatics,* **Int. J.** *on Telecomm. & Information Technology,* 11(4), (403-428).

Stilman, B., (1994d) A Linguistic Geometry for Control Systems Design, **Int. J.** *of Computers and Their Applications,* (89-110), Vol. 1, No. 2, Dec. 1994.

Stilman, B., (1995a) Deep Search in Linguistic Geometry, SYMPOSIUM ON LINGUISTIC GEOMETRY AND SEMANTIC CONTROL, *Proc. of the First World Congress on Intelligent Manufacturing: Processes and Systems,* (868-879), Mayaguez, **Puerto Rico**, Feb. 1995.

Stilman, B., (1995b) A Linguistic Geometry for 3D Strategic Planning, *Proc. of the 1995 Goddard Conference on Space Applications of Artificial Intelligence and Emerging Information Technologies,* (pp. 279-295), NASA Goddard Space Flight Center, Greenbelt, MD, **USA**, May 1995.

Stilman, B., (1995c) A Linguistic Geometry for Multiagent Systems, *Proc. of The 8th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems* - IEA/AIE, (3-12), Melbourne, **Australia**, June 1995.

Stilman, B., (1995d) Managing Search Complexity in Linguistic Geometry, **Int. J.**: *IEEE Transactions on Systems, Man, and Cybernetics,* (submitted).

Stilman, B., (1995e) Network Languages for Concurrent Multiagent Systems, **Int. J.** *Computers & Mathematics with Applications.,* (submitted).

Stilman, B., Dyer, D. (1995f) Linguistic Geometry for Aerospace Combat Simulation: Serial and Concurrent Agents, *Proc. of the 5th International Conference on Human-Machine Interaction and Artificial Intelligence in Aerospace* - HMI-AI-AS'95, (invited paper), Toulouse, **France**, Sept. 1995, (to appear).

Stilman, B., (1995g) Linguistic Geometry for Robotic Systems: From Partial to Total Concurrency, *Proc. of the First ECPD International Conference on Advanced Robotics and Intelligent Automation,* (invited paper), Athens, **Greece**, Sept. 1995, (to appear).

Stilman, B., (1995h) Heuristic Networks for Aerospace Systems Control: Totally Concurrent Motions, *Proc. of the Int. Conference on Recent Advances on Mechatronics* - ICRAM'95, Istanbul, **Turkey**, August 1995.

Stilman, B., (1995i) Linguistic Geometry: A Formal Language for Solving Multiagent Discrete Differential Games, *Proc. of the* IEEE/ISIC *Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems,* (invited paper), Monterey, CA, **USA**, August 1995, (to appear).

Stilman, B., (1995j) Total Concurrency and Uncertainty in Linguistic Geometry, *Proc. of the IEEE International Symposium on Intelligent Control,* Monterey, CA, **USA**, August 1995 (to appear).

Yakhnis, V., Stilman, B., (1995a) Foundations of Linguistic Geometry: Complex Systems and Winning Conditions, SYMPOSIUM ON LINGUISTIC GEOMETRY AND SEMANTIC CONTROL, *Proc. of the First World Congress on Intelligent Manufacturing: Processes and Systems,* (843-854), Mayaguez, **Puerto Rico**, Feb. 1995.

Yakhnis, V., Stilman, B., (1995b) A Multi-Agent Graph-Game Approach to Theoretical Foundations of Linguistic Geometry, *Proceedings of the Second World Conference on the Fundamentals of Artificial Intelligence* (WOCFAI 95), Paris, **France**, July 1995.

Inversion of Atmospheric Radiance Measurements for the
Measurement of Temperature, Velocity, and
Turbulence: A Preliminary Study

David W. Watt
Associate Professor of Mechanical Engineering
Kingsbury Hall
The University of New Hampshire
Durham, NH 03824

Final Report for :
Summer Faculty Research Program
Phillips Laboratory, Hanscomb AFB

Sponsored by :
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Phillips Laboratory
Hanscomb AFB, MA

September, 1995

Inversion of Atmospheric Radiance Measurements for the Measurement of Temperature, Velocity, and Turbulence: A Preliminary Study

David W. Watt
Associate Professor of Mechanical Engineering
Kingsbury Hall
The University of New Hampshire
Durham, NH   03824

Abstract

The feasibility of multi-spectral imaging of atmospheric radiance for remotely detecting turbulence and wind fields was studied. A model for fluctuations in atmospheric radiance is proposed, and an approach to inversion of atmospheric radiance data to obtain temperature, turbulence structure parameter, and velocity profiles is discussed. The approach is based on the assumption of locally isotropic turbulent mixing, which gives the instantaneous temperature field a predictable statistical structure. A preliminary assessment of the feasibility of the technique is presented and prospects for further research are discussed.

Inversion of Atmospheric Radiance Measurements for the Measurement of Temperature, Velocity, and Turbulence: A Preliminary Study

David W. Watt

1. Introduction

Passively emitted atmospheric radiance in the lower atmosphere (below 30 km) depends on the composition of the gas phase, the concentrations of aerosols, cloud cover and precipitation and the local temperature. The composition of the atmosphere varies dramatically in the vertical direction as the density and temperature diminish and different chemical species are present in different proportions. In the lateral direction, the effects of large weather systems, and topography cause spatial variations in the structure of the atmosphere over lateral scales on the orders of 10 to hundreds of miles. In addition to these large scale variations, there are local structural variation in the atmosphere caused by a variety of dynamic processes commonly grouped together under the name turbulence. Turbulent motion in the atmosphere occurs on scales of millimeters to hundreds of meters; this motion results from the breakdown of large-scale coherent motion in the atmosphere (such as thermal plumes or gravity waves) into randomly oriented vortical motions often referred to as eddies.

The effect of turbulence is to enhance the transfer of heat, mass, and momentum from the surface into the atmosphere and to enhance mixing of aerosols throughout the atmosphere. As a consequence of this activity, the temperature field (and to a lesser extent the chemical concentration fields) take on a spatially random, self-similar structure of a range of scales. As a result of the thermal structural variations, the atmospheric radiance and the local refractive index both display variations characterized by the length scale of the

temperature fluctuations. The effect of the refractive index fluctuations on imaging through atmospheric paths has been the subject of extensive research in attempts to understand stellar scintillation, laser beam degradation, and imaging effects.

The effects of spatial variation of atmospheric radiance has received a different type of attention. A great deal of attention has been paid to the understanding of the physics of atmospheric radiation; the resulting models have been incorporated into the LOWTRAN, MODTRAN, HITRAN, and FASCODE atmospheric radiance codes. These codes are designed to predict the absorptance and radiance with moderate to (very) high spectral resolution and to account for a wide range of atmospheric conditions, scattered sunlight, and other ambient factors. Consequently, these models treat the atmosphere as a relatively small number of layers, each having vastly different properties. Experimentally, atmospheric radiance is measured using filtered radiometers or Fourier Transform Spectrometry (FTS). Both the models and the measurements focus on accurate understanding of the chemistry and on the dynamic effects of long-time scale (diurnal and seasonal) variations.

The use of emitted atmospheric radiance to remotely sense the character of the atmosphere has received very little attention; for example, some previous work in this area has concerned the use of the PL/GPOL codes to invert FTS measurements to obtain vertical temperature profiles (see Smith, et. al.; Theriault, et. al). This report discusses the potential use of multispectral images of passively emitted atmospheric radiance for remotely sensing turbulence, temperature, and velocity fields and begins to address the shortcomings of prior work (Smith and Herter, 1985; Watt and Philbrick, 1991). Among the applications for this technique would be to use imaging radiometers as a mean to measure

refractive structure parameter C_n^2 profiles, provide "ballistic wind"

measurements, and to detect clear air turbulence and wind shear. The theory and

simulations were done by the author at Phillips Laboratory at Hanscomb Air Force

Base as part of the work as an AFOSR Summer Faculty Fellow in the Summer of 1995.

2. A Model For Time Evolution of Atmospheric Radiance Images

The formation of these images can be modeled by considering an imaging

radiometer staring into the atmosphere; the image plane of the radiometer has an

array of detectors, and the detector has an imaging volume defined by the imaging

lens, aperture, and detector. The atmosphere within the imaging volume can be

considered to be a series of parallel slices perpendicular to the view axis; the

composition (and therefore the radiative properties) of the atmosphere can vary

along the view axis, but it should be essentially constant within an individual

slice. The temperature field may also vary significantly along the view axis; in

addition, temperature fluctuations due to turbulence are present within an

individual slice. When the width w of the slice is less than the integral scale

of turbulence, the variance of the temperature fluctuation is shown to be

$<\delta T>^2 \approx C_T^2 w^{2/3}$, where C_T is the well-known structure parameter (see Tartarski). In

imaging applications, the width of the slice grows linearly with the distance z

from the receiver so that $<\delta T>^2 \sim C_T^2 z^{2/3}$.

The effects of both the mean temperature distribution and the small

temperature fluctuations on images of atmospheric radiance are seen by examining

the radiative transfer equation: if we neglect scattering, the intensity I_λ

striking a detector from an imaging volume intermediate atmosphere is obtained by

$$\overline{I_\lambda} = \int_0^\infty J_\lambda(T(z)) \, a_\lambda(z) \exp\left(-\int_0^z a_\lambda(z')dz'\right)dz \tag{1}$$

where J_λ is the spectral black body intensity, a_λ is the spectral absorptivity, z is the distance down the optic axis, and T is the temperature. In imaging radiometers, the variation of intensity across the detector is determined by the lateral variation in the temperature. Assuming that the radiometer has an imaging lens of focal length F with the detector in the focal plane, the intensity distribution $I(x_I, y_I)$ in the image plane is given by

$$I_\lambda(x_I, y_I) = \int_0^\infty J_\lambda(T(x_{Iz}/F, y_{Iz}/F, z)) \, a_\lambda(z) \exp\left(-\int_0^z a_\lambda(z')dz'\right)dz \tag{2}$$

where (x_I, y_I) are the image plane coordinates. Note that the factor z/F accounts for the magnification. Defocussing effects are ignored in this application. In order to relate the spatial temperature variation δT to the intensity variation δI we assume that the temperature fluctuations are small relative to the mean temperature and linearize the emitted intensity to obtain

$$\delta I(x,y) = \int_0^\infty \frac{\partial J_\lambda}{\partial T}\Bigg|_{\overline{T}(z)} \delta T(x,y,z) \, a_\lambda(z) \exp\left(-\int_0^z a_\lambda(z')dz'\right)dz \tag{3}$$

If we assume that the width of the slice is less than the integral scale of turbulence, we can use the structure parameter to estimate the temperature fluctuations, the relation between the structure parameter distribution and the variance of the radiance fluctuations is given by

$$<\delta I_\lambda^2> \approx \int_0^\infty C_T^2(z) \, z^{2/3} \left[\frac{\partial J_\lambda}{\partial T}\Bigg|_{\overline{T}(z)} a_\lambda(z) \exp\left(-\int_0^z a_\lambda(z')dz'\right)\right]^2 dz \tag{4}$$

Equations (2) and (3) would provide the basis for obtaining profiles along

39-6

the optic axis of the temperature and structure parameters if: 1) there is enough spectral and spatial variation in the spectral absorptivity, 2) a large enough number of spectral intensity and variance measurements are made; and 3) an accurate model of the spectral absorptivity is available.

The effect of the ambient fluid velocity field for short time intervals can be examined in the context of the "frozen turbulence" assumption; that is, the temperature field deforms slowly with respect to the overall velocity. Therefore, we assume that over a short time interval t, the temperature fluctuations evolve as $\delta T(x,y,z,t+\tau) = \delta T(x+U\tau,y+V\tau,z,t)$, where U and V are the velocities normal to the optic axis. In addition to the frozen turbulence assumption, we further assume that the motion parallel to the optic axis results in a negligible magnification change. Based on these assumptions, the short-time evolution of the radiance images depends only on the motion normal to the optic axis. The effect of the velocity components can be seen explicitly by examining the cross-correlation of two radiance images, I_1 and I_2, separated in time by a small interval. The cross-correlation function $R(u,v)$ is given

$$R_{II}(U_i, V_i) = \iint I_1(x_i, y_i) \, I_2(x_i + U_i, y_i + V_i) \, dx \, dy \qquad (5)$$

by where U and V are the spatial shifts in the detector plane. The velocity at any location z along the optic axis has an apparent magnitude when viewed in the image plane that is scaled by the magnification F/z; that is, the displacement U of an object at a location z will move across the detector by an amount F/z. Therefore, the cross-correlation function can be expressed as

$$R_{II}(u_\perp, v_\perp) = \int_0^\infty \int_{-\infty}^\infty \int [\frac{\partial I_\lambda}{\partial T}\bigg|_{\overline{T}(z)} a_\lambda(z)\exp(-\int_0^z a_\lambda(z')dz')]^2 \cdot$$

$$\delta T(x_\perp z/F, y_\perp z/F, z) * \delta T([x_\perp + u_\perp'(z)]z/F, [y_\perp + v_\perp'(z)]z/F], z)\,dxdydz$$

(6)

where $u_\perp(z) = U(z)\tau F/z$ and $v_\perp(z) = V(z)\tau F/z$ are the apparent displacements of an object at z. This equation can be simplified. First, let

$$A(\lambda, z) = [\frac{\partial I_\lambda}{\partial T}\bigg|_{\overline{T}(z)} a_\lambda(z)\exp(-\int_0^z a_\lambda(z')dz')]$$

(7)

Second, note that the spatial inplane cross-correlation of the temperature field can be expressed in terms of the structure function, i.e.

$$R_{TT}(u, v) = \int\int \delta T(x, y)\,\delta T^*(x+u, y+u)\,dxdy$$
$$\approx C_T^2 z^{2/3} c_{TT}(u, v)$$

(8)

where c(u,v) is the correlation coefficient. Therefore, the equation for the image cross-correlation functions takes a form similar to the equation for the intensity variance and is

$$R_{I_{1,\lambda}I_{2,\lambda}}(u_\perp, v_\perp) \approx \int_0^\infty A[\lambda, z]^2 C_T^2 z^{2/3} c_{TT}(u_\perp'(z)\,z/F, v_\perp'(z)\,z/F, z)\,dz$$

(9)

Finally, we see that the image cross-correlation function can be related to the temperature cross-correlations at each slice location z. As in the case of the intensity variance, the correlation functions can be obtained from a sufficiently large set of spectral image cross-correlations. Once the temperature cross-correlations c_{TT} are obtained, the inplane velocity can be found from the displacement of the correlation peak for each location z.

3.0 Radiance Measurements for Velocimetry

In general, atmospheric radiance has not been used as a means for detecting turbulence. Fourier-Transform Spectrometry (FTS) is used to make high spectral resolution radiance measurements (see Smith, et. al, Theriault, et. al.); it is typical for FTS to provide accurate characterization of several thousand vibration-rotation lines. However, FTS obtains average radiance values which are unsuitable for turbulence estimates: first, the fact that FTS collects radiation from a sample volume onto a single detector provides a spatial average of irradiance; second, since FTS requires a long scan time to make the measurements, there is significant time averaging as well.

Multi-spectral imaging, or imaging spectroscopy is a technique whereby simultaneous spectral images are obtained. In general, these images are formed by focussing the radiation collected from the imaging volume onto a diffraction grating; the grating disperses the spectrum so that both spectral and spatial information may be collected. The spatial information is obtained using scanning mirrors in conjunction with line or area arrays. Generally, these instruments have been used in the UV-NIR to provide reflectance spectra. The ultimate extension of this technology is probably the space-borne AVIRIS experiment, which provides coverage of 220 spectral bands in the UV-NIR region from .4 to 2.45 microns. More "common" multispectral instruments may provide coverage of about 10 spectral bands.

In order to implement the ideas presented in section two, images of the atmosphere that are spatially resolved in a number of wavebands must be used. Since the atmosphere radiates most strongly in the infra-red, a system that is sensitive in the spectral range between 5 and 15 microns would probably be required. However, the usefulness of a given waveband depends on the

distribution of the integral weighting function and whether the atmospheric temperature fluctuations produce sufficiently strong radiance fluctuations to produce a useable signal on the detector. These factors were evaluated in a numerical simulation.

3.0 Simulations

Equation (1), (4), and (9) together form a basis for sensing the temperature, structure parameter, and velocity. These equations are coupled in that the mean temperature distribution is required to evaluate the blackbody intensity I_λ and its temperature derivative $\partial I_\lambda / \partial T$ and that the structure parameter distribution is required to obtain the velocity. The optimal solution of these equations was beyond the scope of this project. However, in order to gain some insight into the nature of the radiance images, we carried out a series of simulations to estimate the relative magnitude of the irradiance fluctuations, and mean irradiance values. The simulations used discretized versions of equations (1),(4), and (9): first, for simplicity we obtain discrete versions of the weighting function in equation 1 at discrete locations (with index k) and at discrete wavelengths (with index l) as in

$$ B_{l,k} = a_\lambda (k\Delta z) \exp(- \sum_{k'=0}^{k} a_\lambda (k'\Delta z) \ \Delta z) \tag{10}$$

A similar discrete version of the weighting function for the turbulence structure parameter (see equations (7) and (9)) is obtained as $C_{k,l} = A[\lambda_l, k\Delta z]^2 (k\Delta z)^{2/3}$. The three equations were then discretized as

$$\overline{I_l} = \sum_{k=0}^{N} J_\lambda(\lambda_l, T(k\Delta z)) B_{l,k} \tag{11}$$

$$\langle \delta I \rangle^2{}_l = \sum_{k=0}^{N} C_{l,k}(C_T^2)_k \tag{12}$$

$$R_{u,v,l} = \sum_{k=0}^{N} C_{l,k}(C_T^2)_k C_{u,v,k} \tag{13}$$

where the subscripted quantities are discrete samples of their continuous counterparts appearing in equations (1),(4), and (9).

The mean radiance and variance equations were modeled to determine the magnitude of these quantities in a number of wavebands. The models used the LOWTRAN data base to provide the spectral absorptivity data and a radiosonde profile provided by Dr. Bob Beland of Phillips Lab for the temperature and C_T^2 data. The model used 24 equally spaced spectral bands in the range from 7 to 14 microns; in addition, the model assumes a relatively wide band pass (5 per cent of the wavelength) at each wavelength. so that the absorptivity is integrated over a large number of spectral lines. Since the radiosonde data has a much higher spatial resolution than the LOWTRAN data base, the spectral absorptivity profiles were interpolated to provide the necessary spatial resolution for the model. The temperature fluctuation magnitude at each location z was computed from structure parameter data; a random field based on the fluctuation amplitude was generated with a Gaussian random number generator.

The two cases discussed here are the simulation of the downwelling radiance and of the radiance from a horizontal path at an altitude of 1000 meters. In both cases, the imaging system was assumed to have an f/2 lens with a focal length of 400 millimeters and a detector with 256 element across a 5 millimeter width. Figures 1 and 3 show the radiance weighting functions $B_{k,l}$ as a function

of range (or altitude) for the horizontal and vertical cases for a number of wavelengths between 7 and 14 microns. Figures 2 and 4 show the structure parameter weighting functions $C_{k,1}$ as a function of range.

In the vertical cases, both weighting functions fall off with distance from the ground; several of these wavelengths show secondary peak in the weighting function between 10 and 15 kilometers. The initial fall-off in the weighting function can be attributed to both the Bouger's Law attenuation of low-altitude gases in the atmosphere (primarily CO_2) and the fall-off in concentration of these gases with altitude(fig 5); the secondary peak correlates with the increase in the ozone concentration with altitude that begins at 10 kilometers. The presence of this plateau in the 8.8 and 9.4 micron weighting functions implies that thermal fluctuation at this altitude are sensible from the ground. The horizontal case at 1 km altitude shows a fall-off in both weighting functions governed entirely by Bouger's law attenuation (since the absorptivity is constant along the path). In both cases, the structure parameter weighting function peaks a short distance from the detector. This maximum is due to the competing effects of atmospheric attenuation of the emitted signal and the increase in fluctuation magnitude as the imaging volume widens with distance.

The computations of the radiance magnitude and fluctuations are shown in figure 5. We see that in both cases the fluctuation magnitude is several orders of magnitude less than the mean radiance. Figure 6 shows the fluctuation magnitude in equivalent temperature. While the low altitude horizontal path shows very strong thermal fluctuations (several degrees or more) over a wide spectrum, the vertical path fluctuations peak at about 1 degree or so over a relatively narrow spectrum. Since the sensitivity of commercially available imaging radiometers is typically quoted at .1° K, the selection of wavelength is

very important in the vertical case.

4.0 Discussion

The use of atmospheric radiance data to measure turbulence relies on accurate measurements of spectral radiance at many wavelengths. Although the simulations indicate that both horizontal and vertical paths produce sensible radiance fluctuations, the work on the inversion of equations (1), (4), and (9) was incomplete and therefore inconclusive so far. We can, however, make a number of observations about the nature of this problem and a number of promising techniques that make this approach worthy of future research.

First, since the mean radiance values are several orders of magnitude greater than the fluctuation magnitude, the dynamic range of many detectors is inadequate to measure both. Therefore, some scheme is required to make accurate measurements of the offset provided by the mean radiance. Commercial imaging radiometers use reference sources or A.C. coupling to separate the mean radiance from the fluctuations. In either case, special care must be taken to measure both quantities simultaneously.

Second, we should note that the solutions of (1) and (4) require accurate estimates of the temperature and atmospheric composition profiles; both of these quantities are used to calculate the weighting coefficients. Although some important constituent profiles (e.g. CO_2) remain fairly constant throughout the year (see figure 7), others, such as ozone (figure 8), vary considerably. The sensitivity of the weighting matrices to variation in temperature and concentration profiles is not readily apparent and is an area for future investigation.

Finally, we must consider the technique's spatial resolution and accuracy together with the method for simultaneous solution of equations (1), (4), and

(9). Ideally, we would like obtain spatial resolutions equivalent to that of a radiosonde (see figure 9) which provides measurements every 20 meters over 30 kilometers or so in altitude. Unfortunately, we are probably restricted to about 10 spectral measurements. This begs the question: what is the spatial resolution of this method? From figures 1 through 4, we can see how the spatial resolution can be determined by the range of visibility through the atmosphere; by using wavelengths whose weighting functions decay rapidly, the range is restricted and the resolution over that range enhanced. Further, the spatial resolution depends on the complexity of the field we are trying to measure. In many cases, the use of wavelet or series expansions allows a reasonably accurate representation of a complex object to be obtained using a small number of coefficients.

The optimal selection of a series expansion is generally based on Karhunen-Loeve (K-L) expansion, a technique which requires detailed knowledge of the statistics of the field in question. First, one computes the joint covariances between all points in the field to form the covariance matrix. Next, one finds the eigenvectors of this matrix. The eigenvector form a basis for representing the field in question; however, this series is typically truncated to based on the ranking of the corresponding eigenvalues. In other words, those eigenvectors with large eigenvalues generally contain most of the signal energy, so that a relatively low-order expansion containing only the most significant eigenvectors provides a reasonable representation of the field. The implication, then, is that a relatively high spatial resolution may be achieved with a small number of measurements. K-L decomposition is a promising approach for this problem since there is a large body of experimental measurements of temperature, composition and C_T^2 profiles that have been generated over the years from numerous balloon launches; indeed Phillips Laboratory is about to initiate a new program of

balloon launches for just this purpose.

In addition to the use of prior statistical information, a number of techniques exist for solving ill-posed, non-linear, inverse problems such as the one described here. These are based chiefly on the imposition of global constraints such as smoothness, energy, and entropy and are well known in the signals processing literature (see Stark, e.g.).

5.0 Future Work

The current study will serve as the basis for summer research extension proposal; this proposal will treat the following issues:

1) Determining the optimal wavelengths for applying this technique in both vertical and horizontal paths.

2) Determining the detectability and other effects of variation in the atmospheric temperature, ozone, and humidity profiles and the effects of scattering and blurring on the images.

3) Assembling an ensemble of radiosonde datato assess the suitability of K-L decomposition to the solution of equations (1),(4),and (7).

4) Assessing by numerical simulation various solution methods of the linearized and nonlinear verision of equations (1),(4), and (7).

5) Design of an appropriate multispectral imager for this application.
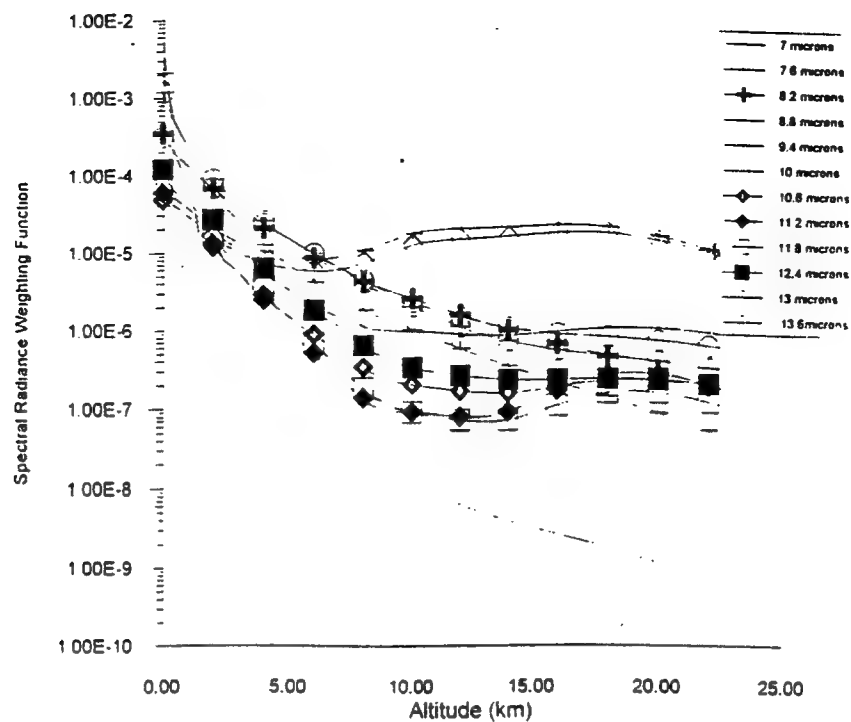
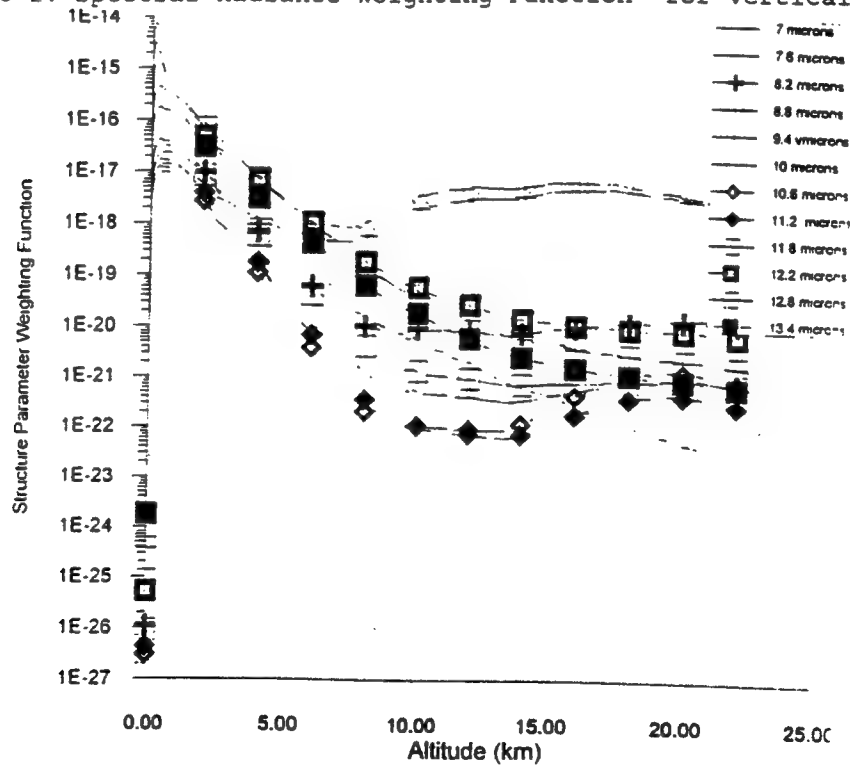Figure 1: Spectral Radiance Weighting Function for vertical path.

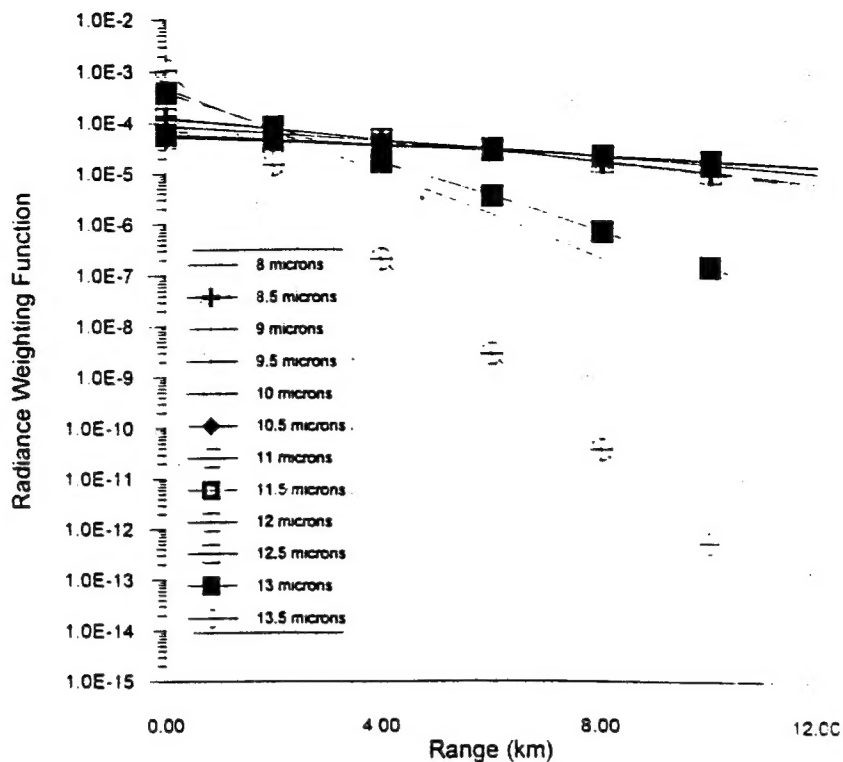Figure 2: Spectral Structure Parameter Weighting Function for vertical path.

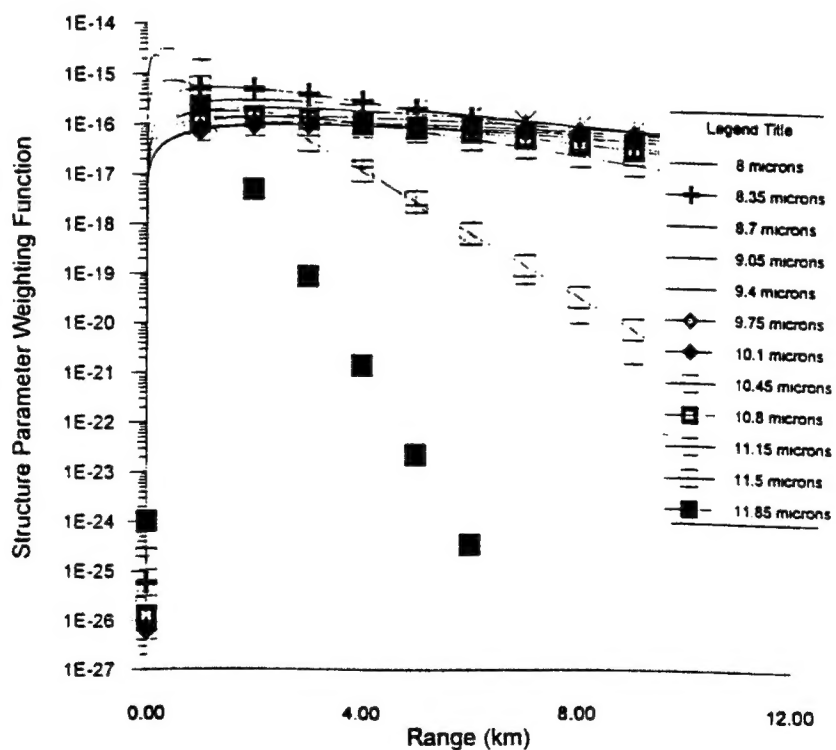Figure 3: Spectral Radiance Weighting Function for horizontal path.

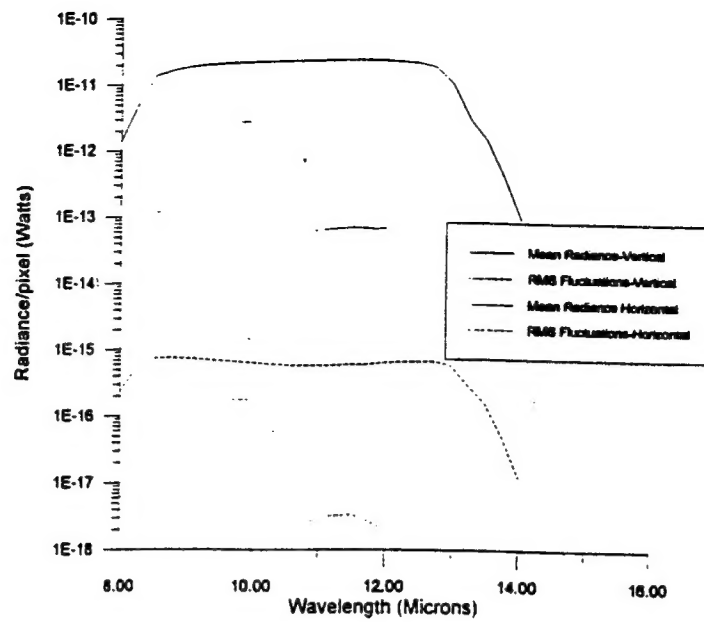Figure 4: Spectral Structure Parameter Weighting Function for horizontal path.

39-17

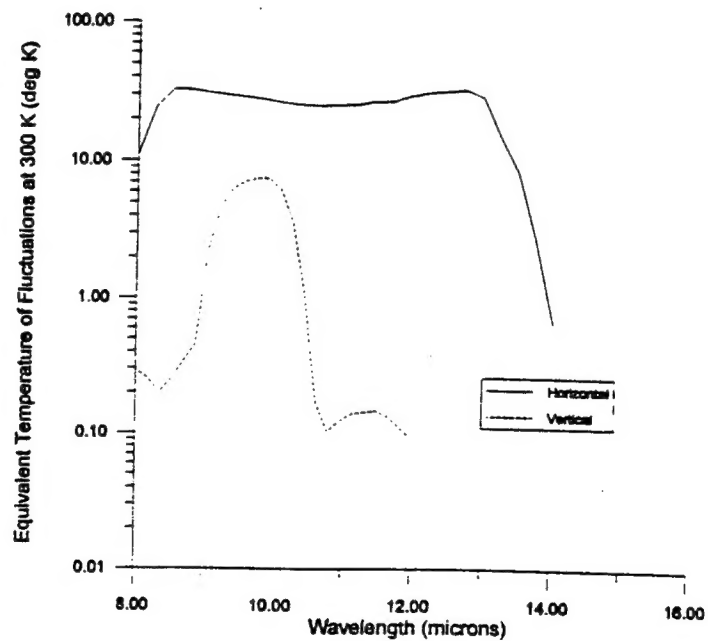Figure 5.   Mean and RMS Radiance for horizontal and vertical paths

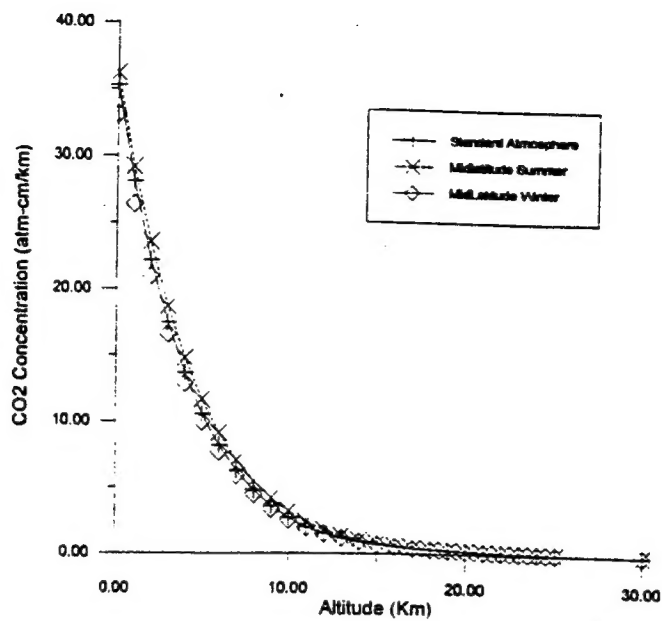Figure 6.   RMS Equivalent Temperature for horizontal and vertical paths

39-18

Figure 7. Seasonal variability of CO_2 Profiles from LOWTRAN database.
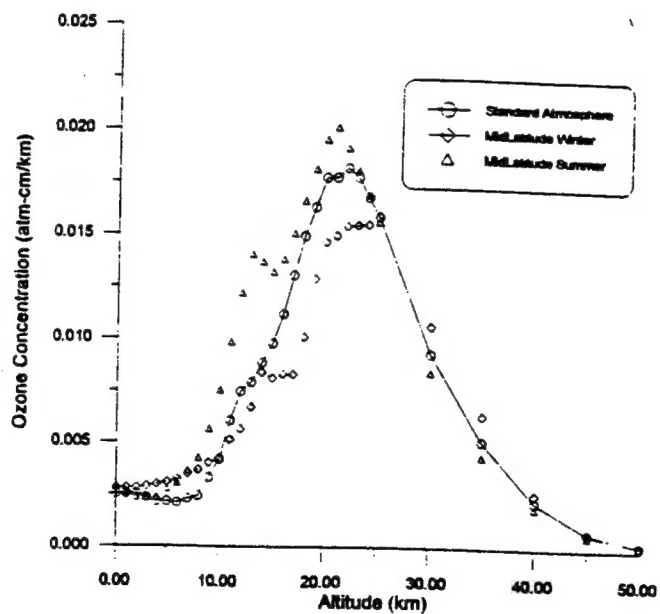

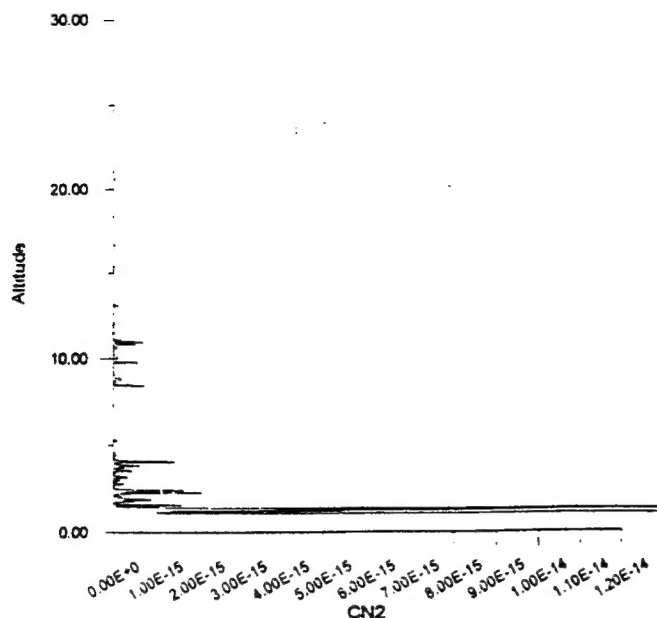Figure 8. Seasonal variability of Ozone Profiles from LOWTRAN database.

Figure 9. C_T^2 Profiles from radiosonde data.

References

Toselli, F. and Bodechtel, J. <u>Imaging Spectroscopy: Fundamental and Prospective Applications</u>,Kluwer, Dordrecht, The Nethelands, 1989.

Smith, L. and Herter, T. "Model for predicting infrared atmospheric emission fluctuations ("sky noise") SPIE PROC. Vol 551, 1985.

Smith, W.L. , Woolf, H.M., and Revercomb, H.E., 1991, Linear Simultaneous Solution for the Retrieval of Temperature and Absorbing Constituents from Radiance Spectra, Appl. Opt. 30, pp117-1123, 1991.

Stark, H. ed., <u>Image Restoration</u>, Academic Press, 1988.

Tatarski,V.I. <u>Wave Propagation in a Turbulent Medium.</u> McGraw-Hill, 1961.

Theriault, J.M, Anderson, G.P, Chetwynd, J.H., Qu,Y. Murphy,E., Turner, V., Cloutier,M, and Smith, A., "Retrieval of Tropospheric Profiles form IR Emission Spectra: Investigation with The Double Beam Interferometric Sounder, 18th Annual Review Conference on Atmospheric Transmission Model, Phillips Laboratory, Hanscomb AFB, June 7, 1995.

Watt, D.W. and Philbrick, D.A. "Images of Turbulent, Absorbing-Emitting Atmospheres and Their Application to Windshear Detection, SPIE Vol. 1467, 1991.